CALIBRATGE: CORRECCIÓ DELS ERRORS DE MESURA I ESTANDARDITZACIÓ DE LA DIETA.

Autor: Guillem Pera Blanco

Director: Erik Cobo

Data: Desembre de 2004

ÍNDEX

1. RESUM

Objectiu: L'error de mesura en la dieta pot provocar l'atenuació dels efectes observats en relacionar aquella amb la incidència de càncer. Aquest projecte pretén descriure, justificar i aplicar mètodes de calibratge usant regressió lineal, per corregir els errors de mesura i estandarditzar les dades provinents de diversos centres, relatives al consum de carn en una cohort europea de més de mig milió de persones de 10 països i la relació amb l'aparició de càncer gàstric (CG).

Mètodes: El calibratge consisteix a aprofitar les dades d'un qüestionari de dieta no esbiaixat (R) aplicat a una part de la cohort, per corregir els *hazard ratios* (HR) del consum de carn amb CG basats en les dades aportades per un mètode de mesura de menys qualitat (Q) però aplicat a tota la cohort. S'obtenen valors predits (calibrats) per a tota la cohort a partir de la regressió entre *R* i *Q*. Aquests valors calibrats s'apliquen a un model de Cox i se n'obtenen estimacions corregides dels HR's. El projecte usarà les dades de l'estudi EPIC.

Resultats: Els coeficients de calibratge varien entre 0,27 i 0,73 , segons el país i el gènere. L'ajust del model lineal entre consum de carn mesurat amb *R* i amb *Q* és molt dolent a nivell individual (r^2<0,20) degut a què s'usa només un record de 24 hores com a mesura de referència, però a nivell grupal r^2 puja fins al voltant de 0,60. L'aplicació de transformacions a les dades o l'exclusió de grans consumidors no aporta millores en l'ajust del model. El HR de CG obtingut per 100 grams de consum de carn es corregeix d'1,43 (IC95%=1,13-1,81) usant les dades originals a 1,97 (IC95%=1,21-3,22) usant les dades calibrades.

Conclusió: Fins i tot si es vulneren algunes assumpcions en què es basa el mètode de calibratge (incorrelació entre els errors de *Q* i *R*, estimacions no esbiaixades del consum real a partir del qüestionari de referència), el calibratge serveix per corregir almenys una part del biaix amb què estimaríem els HR's si uséssim només les dades del qüestionari general. En definitiva, el calibratge és un recurs per disminuir els efectes de l'error de mesura de la dieta en l'estimació dels paràmetres d'associació d'aquella amb la malaltia.

2. ESTRUCTURA DEL PROJECTE

El projecte s'estructura en els apartats habituals d'un article científic: *Introducció, Població i materials, Mètodes estadístics, Resultats, Discussió* i *Conclusió,* a més de la portada, l'índex, les referències bibliogràfiques i els agraïments. També s'inclou un anex on es podrà consultar un *Índex d'abreviatures i símbols*, una extensió més detallada d'una part dels resultats (*Modificacions al model de calibratge original*) no indispensable per al seguiment del projecte que s'ha posat a l'anex per no fer tan extens l'apartat de *Resultats*, un *Esborrany per a un article* que és una versió resumida de tot el projecte (les referències, taules i figures s'han de buscar en el cos del projecte) i el codi dels programes informàtics usats per obtenir els resultats. A més, s'inclou un *Resum* que permet una visió ràpida dels objectius, mètodes, resultats i conclusions del projecte.

El projecte es basa en una part de la feina que desenvolupo a l'Institut Català d'Oncologia. Per tant he aprofitat part d'aquesta feina, com l'empresa també es beneficia dels coneixements adquirits a mesura que anava desenvolupant el projecte. La quantificació d'hores en que he treballat sobre el calibratge en l'àmbit laboral és gairebé impossible, ja que la primera vegada que vaig estudiar aquests mètodes fou el 1996. La redacció del projecte en si i l'execució dels primers programes específics començà a mitjans de març de 2004, si bé des d'un mes abans ja em vaig dedicar a examinar la bibliografia que encara no coneixia. Per diverses raons vaig aturar el projecte el mes de maig, per reemprendre'l el setembre de 2004, i l'he acabat a finals de novembre. La dedicació mitjana ha estat d'unes 16 hores setmanals durant 24 setmanes, amb una càrrega resultant d'unes 380 hores (taula 1). De la previsió d'hores que havia de dedicar al projecte només m'he desviat clarament en l'apartat de revisió bibliogràfica, amb més de 100 cites revisades. Cal remarcar que molts dels coneixements adquirits a mesura que anava recopilant bibliografia o fent els anàlisis no s'han presentat en aquest projecte per manca d'espai, però evidentment es veuran reflectits en la meva tasca laboral.

Taula 1. Càrrega d'hores dedicades a cada part del projecte.

Tasca	**Hores**
Revisió bibliogràfica	150
Anàlisi i manipulació de dades	140
Redacció	60
Revisió	20
Preparació presentació	10

Per últim, una consideració lingüística: fins poc dies abans d'acabar aquest projecte m'he estat referint al procés de calibratge com a procés de *calibració*. Aquest terme és incorrecte, segons consta en el Diccionari de la Qualitat (Cisneros 2001) que recull termes tècnics, inclosos els termes estadístics, en català, alemany, anglès, castellà i francès. La paraula *calibració* en català i *calibración* en castellà és una traducció directa de la paraula anglesa *calibration*. Els termes correctes són *calibratge* i *calibrado* en català i castellà respectivament.

3. AGRAÏMENTS

Abans de res, vull agrair a l'Erik Cobo, professor de la FME de la UPC i director d'aquest projecte, els seus consells i comentaris.

Donar les gràcies també als meus co-tutors del Servei d'Epidemiologia i Registre del Càncer (SERC) de l'Institut Català d'Oncologia (ICO), el Víctor Moreno, professor titular de Medicina Preventiva de la UAB, i el Carlos Alberto González, coordinador de l'estudi EPIC a Espanya i coordinador europeu de l'EUR-GAST, pel seu generós suport i ajut.

A la resta dels companys del SERC agrair la seva companyonia, especialment a la Mireia Díaz, que em va engrescar a fer la llicenciatura i em va fer de guia en aquesta carrera, i també al Toni Berenguer, Ramon Clèries, Àngela Twose, Gina Albero, Toni Agudo i Juan Ramón González (SPCC) per la seva ajuda al llarg de la carrera i en el projecte.

Als companys de classe, sobretot a la Sara Giner i la Laura Muñoz, "proveïdores oficials d'apunts", i a la Núria Pérez per les seves galetes (que em cruspia) entre classe i classe.

Al Nick Day (Cambridge), Pietro Ferrari (Lió) i Heindrek Boshuizen (Amsterdam), col·legues estadístics de l'EPIC Europa, pels seus brillants consells.

Al Josep Maria Pera, el meu pare, per fer la correcció del català d'aquest projecte, i sobretot per animar-me sempre a continuar els estudis.

I un agraïment molt especial a la Raquel, per la paciència i comprensió en aquests anys en què no ha estat fàcil combinar estudis, feina i família.

A tots vosaltres, MOLTES GRÀCIES!.

4. INTRODUCCIÓ

L'error de mesura és un dels problemes que sempre ha acompanyat la Ciència. Gairebé cap estudi se'n pot lliurar i, com a màxim, es pot intentar minimitzar-lo perquè la influència sobre les conclusions derivades de l'experiment sigui negligible. L'error de mesura i el control d'aquest també és un dels reptes principals als quals s'enfronta l'Estadística. La qualitat de les mesures és fonamental en qualsevol àmbit, però pren un especial interès en el camp de les Ciències de la Salut, on contínuament es prenen decisions basades en mesuraments (White 2003).

En aquest projecte em centraré en l'error de mesura de la dieta, portada a terme amb diversos instruments i en diferents poblacions, i en com es pot intentar corregir aquest error quan relacionem la dieta habitual d'un individu amb l'aparició d'una malaltia rara i crònica com és el càncer.

Una de les eines més utilitzades per corregir l'error de mesura de la dieta quan calculem un paràmetre que relacioni dieta i malaltia és el calibratge (Stürmer 2002). A mitjans dels 80, Armstrong (1985) fou el primer en usar el calibratge per a models lineals generals per corregir l'error de mesura, tot i que és a Rosner (1989), que l'aplicà després a models logístics, a qui molts n'atribueixen la creació.

En un estudi etiològic univariant en què volem saber l'efecte d'una certa variable sobre una determinada malaltia, si la variable és mesurada amb error, l'efecte observat és menor que l'efecte real (atenuació de l'efecte). Encara més, si treballem amb un model multivariant i hi ha una o més variables mesurades amb error, l'efecte observat pot estar tan sobre com infraestimat, fins i tot pels efectes de les variables mesurades correctament (ja que podrien estar correlacionades amb les variables mesurades amb error) (Greenland 1980, Kupper 1984).

A més, en grans estudis multicèntrics, en què els qüestionaris de dieta acostumen a ser diferents entre centres per capturar les dietes locals (Friedenreich 1994), la magnitud i la natura dels errors sistemàtics i aleatoris pot variar entre els centres i distorsionar

l’estimació i interpretació de la relació global entre la dieta i la malaltia quan es combinen les diferents cohorts (Slimani 2002).

Una de les eines més utilitzades per corregir l’error de mesura de la dieta quan calculem un paràmetre que relacioni dieta i malaltia és el calibratge (Stürmer 2002).

El primer objectiu del calibratge és, a nivell individual, intentar corregir el biaix d’atenuació en el risc relatiu (RR) (o altres mesures d’associació) degut als errors aleatoris de la mesura de la dieta (Slimani 2002). En estudis multicèntrics el segon objectiu seria, a nivell ecològic, ajustar per una sobre o infraestimació sistemàtica de la dieta a cada centre.

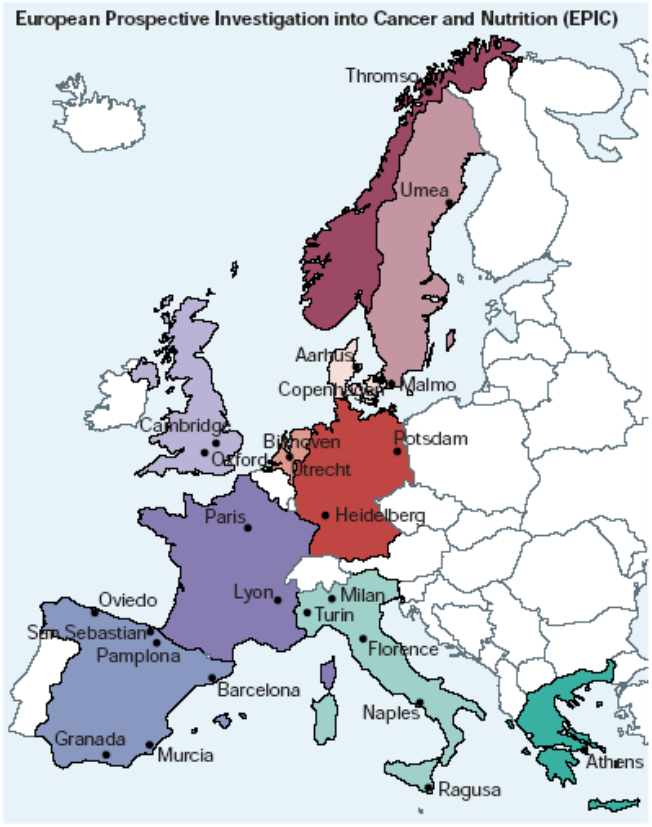
Aquest projecte pretén descriure, justificar i aplicar mètodes de calibratge usant la regressió lineal, per corregir els errors de mesura i estandarditzar les dades provinents de diversos centres relatives al consum de carn en una cohort europea de més de mig milió de persones de 10 països i la relació amb l’aparició de càncer gàstric (CG), avaluada mitjançant un model de riscos proporcionals de Cox.

5. POBLACIÓ I MATERIALS

5.1 L'ESTUDI EPIC

L'estudi EPIC (*European Prospective Investigation into Cancer and nutrition*) és un estudi prospectiu multicèntric sobre dieta i càncer que inclou 28 cohorts de 10 països d'Europa Occidental (França, Itàlia, Espanya, Regne Unit, Alemanya, Holanda, Grècia, Suècia, Dinamarca i Noruega) (Figura 1).

Figura 1. Distribució geogràfica dels centres participants a l'estudi EPIC.



Les característiques més importants de l'EPIC són la grandària (521.468 persones), la distribució geogràfica i l'heterogeneïtat dels patrons dietètics, factors socioculturals i hàbits de vida dels seus participants. Es disposa d'informació sobre la dieta habitual, estil de vida, factors ambientals i antropometria, així com d'una mostra de sang, tot recollit en el moment del reclutament, efectuat entre març de 1991 i abril de 2000. El

seguiment mitjà és de 5,0 anys (varia entre 1 dia fins a 10,8 anys). La majoria de casos de càncer i morts han estat detectats mitjançant un seguiment passiu, creuant bases de dades amb registres de tumors, hospitalaris i de mortalitat. Alguns països han adoptat un seguiment actiu, contactant directament amb els participants o els seus familiars per detectar els casos. La informació sobre la dieta habitual (referida a l'any anterior) es va recollir a través de qüestionaris de freqüencia alimentària (QFA) o història de dieta (HD) desenvolupats i validats a cada país. A més, es va administrar un record de 24 hores (R24H) a 36.994 persones per ser usat com a mètode de referència del calibratge. Aquest mètode usava un *software* creat *ad-hoc* (EPIC-SOFT) que permetia estandarditzar al màxim el consum alimentari entre els diversos centres i incrementar la versemblança de que els errors de mesura fossin similars entre els centres (Slimani 2002).

La població reclutada per l'EPIC no va ser escollida perquè fos representativa de la població general. Es va primar més una alta participació i la seguretat d'un bon seguiment (Riboli 2002). Molts sí que eren reclutats de la població general, però no sempre (participants en cribatge de càncer de mama (Utrecht, Florència), mestres (França), vegetarians i gent preocupada per la seva salut (Oxford) i donants de sang (Itàlia i Espanya)). A França, Noruega, Utrecht i Nàpols només es van reclutar dones (Slimani 2002). L'edat mitjana de la cohort al moment del reclutament era de 51,7 anys (rang 16-98) tot i que un 90% de la mostra tenia entre 32 i 66 anys.

L'estudi específic del CG es desenvolupa dins un subprojecte de l'EPIC: l'EUR-GAST (Estudi sobre Factors Ambientals, Helicobacter Pylori, Susceptibilitat Genètica i Risc de Càncer Gàstric a la Població Europea).

5.2 L'ESTUDI DE CALIBRATGE

La mostra de calibratge es va seleccionar de forma aleatòria de la mostra principal, ponderada per gènere i edat tenint en compte el número de casos esperats en 10 anys de seguiment. L'objectiu era recollir uns 4.000 participants per país. El mostreig també va tenir en compte la distribució per dies de la setmana i estacions de l'any (Slimani 2002). Finalment disposem de 36.994 individus amb R24H.

L'estudi de calibratge es va dur a terme entre el març de 1995 i el juny de 2000. Segons el país, es va trigar entre 10 a 31 mesos en recollir tots els R24H. A França, Alemanya i Dinamarca es van administrar els R24H només 40 minuts després de recollir el QFA. Els països que ja havien fet la major part del reclutament al moment de fer el calibratge van trigar fins a 34 mesos en administrar el R24H desde l'administració del qüestionari basal. L'avantatge dels primers es que asseguren una participació més alta i que la dieta es refereix al mateix període de temps, però a costa d'un major risc de correlació dels errors dels dos qüestionaris. La taxa de participació va variar entre un 54% (Grècia) i un 92% (Alemanya), amb 7 països per sobre del 75%. La duració mitjana de l'entrevista fou de 31,1 minuts i el número d'ítems recollits varia entre 15 (Grècia) i 30 (Regne Unit). Els divendres van ser el dia més difícil per recollir el consum alimentari (algunes entrevistes referides a dissabte es feien els dilluns). Anàlogament, els mesos d'estiu també varen estar infrarepresentats (Slimani 2002).

Per mirar la representativitat de la mostra de calibratge es van comparar les variables més rellevants respecte a la mostra principal. No es van trobar diferències (després d'ajustar per edat) per talla, pes, índex de massa corporal (IMC) i hàbit tabàquic. Es van trobar lleus diferències per nivell educacional i activitat laboral (menys individus sense estudis i desocupats a la mostra de calibratge). També es va mirar si hi havia diferències respecte al consum QFA/HD dels 16 grans grups alimentaris estudiats en cada un dels 28 centres. En un 89% de les combinacions centre-grup alimentari no es van trobar diferències majors del 10% (Slimani 2002).

5.3 MÈTODES DE MESURA DE LA DIETA EN L'EPIC

5.3.1 EL MÈTODE DE REFERÈNCIA

El mètode que s'usa com a qüestionari dietètic de referència és el R24H, un dels més usats com a qüestionari de referència. En l'EPIC s'efectua una sola administració del R24H. Com el nom indica, l'enquestat ha de reportar tot allò que ha menjat durant un dia (habitualment el dia anterior a l'entrevista). D'aquesta forma el biaix de memòria es redueix, en ser el consum molt recent. Un sol qüestionari de R24H no és representatiu de la dieta habitual a nivell individual, per ser referit a un sol dia, però sí a nivell grupal.

El R24H es considera ideal per a comparacions interculturals del nivell de consum mitjà, i és essencialment un mètode obert que permet una descripció força detallada d'un gran nombre de plats i receptes heterogenis (Witschi 1990). Comparat amb els diaris dietètics, els avantatges del R24H són taxes de participació més altes i que l'entrevistador pot supervisar la recollida d'informació i resoldre dubtes (Kaaks 1997).

Un dels problemes d'aquest mètode, com d'uns altres en què intervé la memòria del subjecte, és l'oblit d'aliments consumits o reportar aliments no consumits (Karvetti 1985). A més, factors com l'edat, el nivell educatiu o l'IMC poden influir en les respostes (Karvetti 1985, Klesges 1995). La percepció de "dietes saludables" fa que alguns individus infrareportin la "mala dieta" i supraestimin la "bona dieta" (Madden 1976).

En usar el R24H (o altres mesures relacionades amb la ingesta recent (com un biomarcador)) introduim una variació addicional (incorporada a l'error) respecte a les mesures del QFA/HD que són a llarg termini (Schatzkin 2003).

En l'EPIC el R24H s'administrà mitjançant un programa informàtic *ad-hoc* anomenat EPIC-SOFT (Slimani 1999). El qüestionari es respon en dues etapes. A la primera s'interroga el subjecte sobre què ha menjat el dia anterior en cada àpat (el dia s'estructura en 11 possibles ocasions d'ingesta), sense massa detall (p.ex.: "vedella"). Un cop es té la llista d'allò que ha menjat en tot el dia es repassa ítem per ítem, detallant

quins aliments, quantitats i formes de preparació. Es pregunta pels aliments més comuns si no han estat descrits pel subjecte (per exemple "pa"). El programa incorpora un control de qualitat en què es comproven valors mancants o extrems. El programa porta incorporat una llista d'aliments per cada país. Si l'individu reporta un nou aliment aquest es grava, però a posteriori es decideix si s'equipara a un aliment predefinit o s'incorpora una nova entrada a la base de dades. Quan l'individu no és capaç de descriure prou bé l'aliment (p. ex. "peix") es registra un genèric (això impedeix que l'entrevistador prengui decisions arbitràries). Els individus poden reportar receptes, aliments aïllats o acompanyants (p. ex: "sucre"). Es van predefinir més de 350 receptes com a receptes estàndard en cas que l'individu no sabés descriure-les (p. ex. "paella de marisc"). Per ajudar a descriure la quantitat i el grau de cocció s'usen fotografies. Tot el procés és estandarditzat per a tots els països. De mitjana es triguen uns 30 minuts per contestar el R24H i, en general, no es van apreciar diferències en l'energia calculada a partir dels R24H per entrevistador (Slimani 2000).

5.3.2 EL MÈTODE GENERAL

La majoria dels centres de l'EPIC van utilitzar diferents qüestionaris per mesurar la dieta habitual a l'inici de l'estudi. Bàsicament, però, els podem agrupar en QFA i HD.

Ambdós mètodes avaluen el consum habitual d'una persona durant un període llarg de temps (en aquest cas un any). El QFA és relativament barat i fàcil d'administrar (en molts estudis s'ha enviat per correu) i és un dels mètodes de mesura de la dieta més comunament usats (Kipnis 2002). Bàsicament consisteix en una graella, més o menys complexa. A les files hi ha una llista d'aliments o receptes estàndard. A les columnes, un indicador de freqüència (per exemple quants cops a l'any, al mes, a la setmana o al dia pren un determinat aliment). L'individu ha d'anar omplint la graella amb els aliments que consumeix indicant-ne la freqüència. Alguns qüestionaris incorporen fotografies, diverses mesures o fins i tot files en blanc per afegir-hi aliments no previstos. En una comparació amb el R24H i mètodes bioquímics, els QFA tenien un infrareport del 30-40% i els R24H un 10-20% per energia (Schatzkin 2003).

El mètode HD s'acostuma a contestar amb l'ajuda d'un entrevistador i un programa informàtic. Es repassen totes les ocasions d'ingesta al llarg del dia (esmorzar, dinar, sopar,...) i es pregunta pel consum habitual al llarg de l'any (en una setmana tipus) en cada una d'aquestes ocasions d'ingesta. Es tenen en compte factors estacionals (per exemple es pregunta "quants mesos a l'any menja aquest aliment?") , tipus de cocció, classe de greixos afegits en cuinar o a les amanides per exemple, i s'utilitzen fotografies per ajudar en les respostes. Hi ha un joc de receptes predefinides, modificables per l'entrevistat. Aquest, a més, pot afegir nous aliments i receptes.

6. MÈTODES ESTADÍSTICS

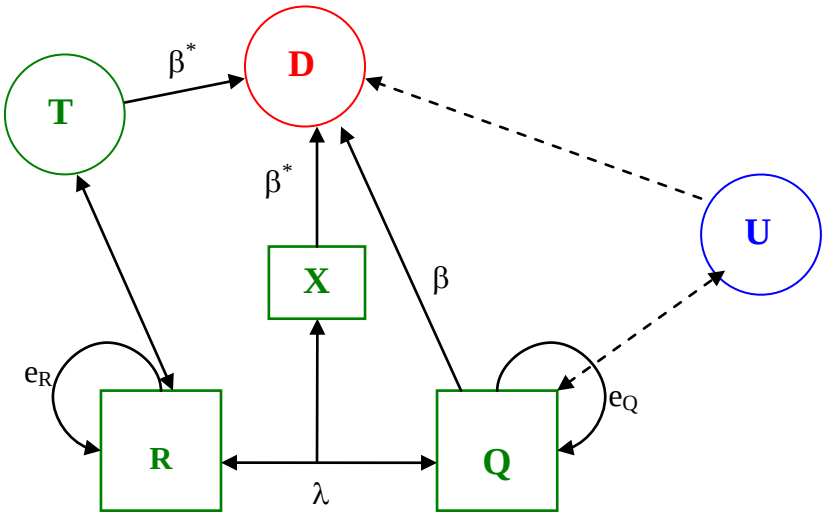
6.1 FONAMENTS TEÒRICS DEL CALIBRATGE

Bàsicament el calibratge consisteix a aprofitar les dades d'un qüestionari no esbiaixat aplicat a una part de la mostra, per corregir els riscos relatius (o altres mesures d'associació) basats en les dades aportades per un (o diversos) qüestionaris de menys qualitat, afectes d'error, però aplicats a tota la mostra. El qüestionari general és més fàcil d'administrar i barat que el qüestionari de referència.

En el diagrama de la figura 2 podem veure esquemàticament què fa el calibratge. Si l'aparició d'una certa malaltia (D) es causada pel consum d'un determinat aliment (T) podem estimar l'efecte d'aquest consum sobre la malaltia (β^*). Se suposa que aquest efecte és estimat assumint que *T* ve mesurat sense error, però habitualment la dieta es mesura amb error. Podem fer servir una mesura dietètica de referència (R) que, si bé té error (e_R), almenys és una estimació no esbiaixada de *T*. Com que aquestes mesures acostumen a ser cares no es poden aplicar a mostres grans. Si necessitem mostres grans (per exemple per detectar efectes lleus) aleshores haurem d'usar un qüestionari de menor qualitat (Q), és a dir, amb més error (e_Q). *R* i *Q* estan relacionades per un coeficient λ (coeficient de calibratge). Podem corregir (calibrar) els valors de *Q* a partir d'*R,* i obtindrem uns valors predits (calibrats) *X* que, en relacionar-los amb la malaltia, mostren l'efecte β^* que hauríem obtingut usant *T*. D'aquesta forma corregim l'efecte atenuat β que obtindríem si uséssim directament *Q* al estimar el risc de malaltia. El model permet la inclusió de variables d'ajust (*U*).

Dins dels fonaments teòrics del calibratge veurem el model en que s'usa directament la variable real per calibrar (apartat 6.1.1), el model que usa la variable de referència *R* (apartat 6.1.2) , l'aplicació específica a un disseny de cohort (apartat 6.1.3) i els avantatges del calibratge en estudis multicèntrics (apartat 6.1.4).

Figura 2. Esquema d'un procés de calibratge. D=malaltia, T=consum real, R=consum de referència (mesura cara i no esbiaixada respecte a *T*), Q=consum mesurat amb error, U=variables d'ajust, X=variable calibrada, β*=efecte real del consum sobre la malaltia, β=efecte del consum sobre la malaltia atenuat per l'error de mesura, λ=pendent de la regressió entre *R* i *Q*, e_R=error del mètode de referència, e_Q=error del mètode general.



6.1.1 EL CALIBRATGE. EL PRIMER MODEL DE ROSNER.

Inicialment Rosner (1990) proposà un model que corregia els errors, tan sistemàtics com aleatoris, que incorporés la variabilitat aportada per la correcció d'aquests errors (suposa un engrandiment dels intervals de confiança; de fet, a més error més engrandiment) aplicat a un model de regressió logística. El model és pensat per a estudis de cohort sobre malalties rares, en que podem comparar els casos amb una gran quantitat de controls, i en què la variable mesurada amb error és contínua.

Suposem que volem estudiar la relació d'una certa malaltia (*D*, variable dicotòmica) amb una sèrie de variables. Lamentablement algunes (*T*, vector de mida k_1x1) són difícils d'obtenir i, per tant, hem d'usar variables indirectes (*surrogates*) més fàcils d'aconseguir però mesurades amb error (*Q*, vector de mida k_1x1). Suposem que la resta de variables (*U*, vector de mida k_2x1) són mesurades sense error. Per il·lustrar aquest model, podem pensar en *D* com el fet de ser diagnosticat o no de càncer durant el seguiment de la cohort, en *T* com la ingesta real habitual d'un cert aliment o nutrient

(mesurada en grams/dia), *Q* la ingesta reportada d'aquest aliment per cada individu, mitjançant un qüestionari aplicat a l'inici de l'estudi, i *U* variables ben conegudes o amb un error de mesura inapreciable com gènere, edat o talla. Més endavant veurem que de vegades és impossible obtenir la mesura real *T* d'una variable. Aleshores ens haurem de conformar amb usar la millor variable disponible (variable de referència *R*) que doni una mesura no esbiaixada de *T*.

El model que volem estimar serà de la forma

$$\text{logit}[\Pr(D \mid T, U)] = \alpha^* + \beta_1^* T + \beta_2^* U \qquad (1).$$

Però com que no disposem de *T* per a tota la cohort hem de fer

$$\text{logit}[\Pr(D \mid Q, U)] = \alpha + \beta_1 Q + \beta_2 U \qquad (2)$$

Immediatament podem obtenir els estimadors esbiaixats (degut a l'error de mesura en *Q*) α, β_1 i β_2 (amb β_1 i β_2 vectors de mida 1xk_1 i 1xk_2 respectivament) a partir de (2) que mostren la relació entre la malaltia i les variables estudiades. Però preferiríem disposar d'uns coeficients corregits α^*, β_1^* i β_2^* , que tinguessin en compte que *Q* ens ve mesurada amb error i que ens diguessin quina és la relació real de *D* amb *T* i *U* després de controlar per *U* i *T* respectivament.

Suposem que podem obtenir la ingesta real *T* mitjançant un mètode car o de difícil aplicació. Usant una mostra relativament petita (submostra de calibratge), en què es mesura la variable difícil d'obtenir amb els dos mètodes, el que té error (*Q*) i el que no en té (*T*), a més de les variables d'ajust *U* utilitzades en el model (2), si fem una regressió lineal múltiple usant *T* com a variable dependent:

$$T = \alpha' + \lambda_1 Q + \lambda_2 U + e \qquad (3)$$

amb α', λ_1 i λ_2 matrius de mida k_1x1, k_1xk_1 i k_1xk_2 , i e vector de mida k_1x1 distribuït segons una N(0, Σ) (*0* vector mitjana de k_1x1 zeros i Σ matriu de covariàncies k_1xk_1), podem obtenir els estimadors corregits

$$\hat{\beta}^* = \hat{\beta}\hat{\lambda}^{-1} \qquad (4)$$

assumint que la distribució condicional de l'exposició real *T* donats *Q* i *U* és la mateixa a la població d'estudi que a la subpoblació de calibratge i que la distribució de l'exposició al *surrogate Q* donada l'exposició real és la mateixa pels individus malalts com sans [Pr(Q|T,D=1)=Pr(Q|T,D=0)] (o sigui, error no diferencial). λ s'anomena factor de calibratge.

L'equació (4) és fàcil d'obtenir si substituïm a (1) *T* per la part dreta de l'expressió (3):

$$\Pr(D \mid Q,U) \cong \exp[c + \beta_1^* \lambda_1 Q + (\beta_2^* + \beta_1^* \lambda_2)U] \qquad (5)$$

Així, $\hat{\beta}_1$, obtingut de la regressió (2) de *D* amb *Q* i *U* serà un estimador consistent de $\beta_1^* \lambda_1$ i $\hat{\beta}_2$ serà un estimador consistent de $\beta_2^* + \beta_1^* \lambda_2$, amb $\hat{\beta}^* = (\hat{\beta}_1^*, \hat{\beta}_2^*)$, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$, i $\hat{\lambda} = \begin{pmatrix} \hat{\lambda}_1 & \hat{\lambda}_2 \\ 0 & I \end{pmatrix}$ amb *0* matriu de mida k_2xk_1 de zeros i *I* matriu identitat de mida k_2xk_2 i $c = \alpha^* + \beta_1^* \alpha'$.

Per tant, el nou coeficient corregit $\hat{\beta}^*$ té en compte la relació entre l'exposició real T i la mesura aproximada d'aquesta exposició *Q* al estimar la relació entre la malaltia *D* i la dieta *T*, a través d'una variable mesurada amb error *Q*. Un cop hem obtingut el coeficient corregit $\hat{\beta}^*$ cal donar-li uns intervals de confiança. Per això cal calcular prèviament la matriu de variàncies-covariàncies de $\hat{\beta}^*$. Aplicant el mètode delta (Rao 1973) en (4) tenim que

$$Cov(\beta_{j1}^*, \beta_{j2}^*) \cong (A' \Sigma_\beta A)_{j_1, j_2} + \hat{\beta} \Sigma_{A, j_1, j_2} \hat{\beta}' \qquad (6)$$

amb $A = \hat{\lambda}^{-1}$ i Σ_{A, j_1, j_2} la matriu de covariàncies referida als elements de les columnes j_1 i j_2 d'A.

Podem obtenir Σ_{A, j_1, j_2} ja que a partir del mètode delta tenim que

$$Cov(A_{i_1 j_1}, A_{i_2 j_2}) \cong \sum_{r=1}^{k}\sum_{s=1}^{k}\sum_{t=1}^{k}\sum_{u=1}^{k}(\partial A_{i_1 j_1} / \partial \hat{\lambda}_{rs})(\partial A_{i_2 j_2} / \partial \hat{\lambda}_{tu}) Cov(\hat{\lambda}_{rs}, \hat{\lambda}_{tu}) \qquad (7)$$

amb $A_{i_r j_s}$ element $i_r j_s$ de la matriu A.

Com que per qualsevol paràmetre y, les matrius de mida kxk $\partial A / \partial y, \partial \hat{\lambda} / \partial y$ estan relacionades per (Searle 1982)

$$\partial A / \partial y = -A(\partial \hat{\lambda} / \partial y) A \qquad (8)$$

podem reexpressar (7) com

$$Cov(A_{i_1 j_1}, A_{i_2 j_2}) \cong \sum_{r=1}^{k}\sum_{s=1}^{k}\sum_{t=1}^{k}\sum_{u=1}^{k}(A_{i_1 r} A_{s j_1} A_{i_2 t} A_{u j_2}) Cov(\hat{\lambda}_{rs}, \hat{\lambda}_{tu}) \qquad (9)$$

Substituint (9) en (6) obtenim que

$$\Sigma_{\beta^*}(j_1, j_2) = Cov(\hat{\beta}^*_{j_1}, \hat{\beta}^*_{j_2}) = (A' \Sigma_{\beta} A)_{j_1, j_2} + \hat{\beta} \Sigma_{A, j_1, j_2} \hat{\beta}' \quad j_1, j_2 = 1, ..., k \qquad (10)$$

amb $k = k_1 + k_2$, $\hat{\beta}^*_{j_i}$ (i=1,2) el j_i coeficient de regressió real, Σ_{β} és la matriu kxk de variàncies-covariàncies de la $\hat{\beta}$ obtinguda en (2), $A = \hat{\lambda}^{-1}$ i Σ_{A, j_1, j_2} és la matriu de variàncies-covariàncies kxk dels elements de les columnes j_1 i j_2 d'A (equació (9)), obtinguda a partir del subestudi de calibratge usant (3).

Cal notar que per obtenir (10) s'assumeix que $\hat{\beta}$ i $\hat{\lambda}$ són variables aleatòries independents (és a dir, la relació entre la ingesta i la malaltia és independent de l'error de mesura del qüestionari).

Per últim, per obtenir l'interval de confiança podem fer

$$\hat{\beta}_j^* \pm z_{1-\alpha/2}\sqrt{Var(\hat{\beta}_j^*)} \qquad (11)$$

amb $Var(\hat{\beta}_j^*)$ obtinguda de $\Sigma_{\beta^*}(j,j)$ a (10) i $z_{1-\alpha/2}$ el percentil superior $\alpha/2$ d'una distribució N(0,1).

Anàlogament, es pot obtenir un interval de confiança per a l'*odds ratio* (OR) exponenciant l'expressió (11).

Quina quantitat d'error de mesura és necessària perquè la correcció sigui útil? A la taula 2 (Rosner 1990) es pot veure com si $\hat{\lambda}$<0,8 les diferències dels OR calculats amb la variable real i els OR calculats amb la variable *surrogate* (sense corregir) són prou diferents.

Taula 2. *Odds ratios* observats* basats en un *surrogate* de l'exposició per diferents nivells del veritable *odds ratio* i $\hat{\lambda}$.

	***Odds ratio* real**			
$\hat{\lambda}$ †	**1,5**	**2,0**	**3,0**	**5,0**
0,2	1,08	1,15	1,25	1,38
0,3	1,13	1,23	1,39	1,62
0,4	1,18	1,32	1,55	1,90
0,5	1,22	1,41	1,73	2,24
0,6	1,28	1,52	1,93	2,63
0,7	1,33	1,62	2,16	3,09
0,8	1,38	1,74	2,41	3,62
0,9	1,44	1,87	2,69	4,26
1,0	1,50	2,00	3,00	5,00

* $OR_{surrogate} = (OR_{real})^{\hat{\lambda}}$

† Coeficient de regressió entre l'exposició real i el *surrogate* de l'exposició (si la desviació estàndard de totes dues mesures fou la mateixa coincidiria amb el coeficient de correlació).

6.1.2 MODEL GENERAL DE CALIBRATGE

Kaaks (7D) dóna una formulació generalista del procés de calibratge. Suposem que la relació (real) entre una certa malaltia (*D*) i el consum habitual d'un cert aliment (*T*) sigui lineal:

$$E[D|T] = \alpha^* + \beta^* T \qquad (12)$$

Estimarem β^* (el paràmetre que ens diu com afecta la dieta a la probabilitat de desenvolupar la malaltia) usant mètodes de regressió. Obtenim valors aproximats de *T* a partir d'un qüestionari dietètic *Q*. La presència d'errors de mesura en *Q* ens portarà a estimar una β esbiaixada, i de fet també una pèrdua de potència estadística, ja que s'atenua l'efecte. Si definim X=E[T|Q] com el consum real mig donat el consum observat *Q* i assumim que E[D|Q,T]=E[D|T] (o sigui, que l'error de mesura és no diferencial), a partir del teorema de Bayes, Kaaks demostra que:

$$E[D|Q] = E[E[D|T]|Q] = \alpha + \beta X \qquad (13)$$

Aquesta equació és idèntica a l'equació (12), excepte que ara usem una estimació del consum *X* en comptes del consum real *T*. *X* és el consum de carn predit (calibrat) a partir del qüestionari *Q*.

La relació entre les mesures del qüestionari *Q* i el consum predit *X*, i la relació entre *Q* i la malaltia *D* depèn de com assumim la distribució dels valors reals de consum *T* i dels errors del qüestionari de mesura. Molts cops s'assumeix normalitat en *T*, amb mitjana μ_T i variància σ^2_T i relació entre *Q* i *T* lineal:

$$Q=\phi_Q+\delta_Q T+\varepsilon_Q \qquad (14)$$

Aquesta equació és una formulació més detallada de $Q=T+e_Q$ on dividim l'error total (e_Q) en sistemàtic (ϕ_Q, δ_Q) i aleatori (ε_Q)), amb l'error aleatori ε_Q distribuït normalment amb mitjana zero i variància $\sigma^2_{\varepsilon Q}$ independent de *T*. Cal dir que la correlació entre un qüestionari de dieta i el consum habitual real acostuma a ser baixa (<0,7) (Willett 1985),

que implica que almenys la meitat de la variabilitat en la mesura de la ingesta és deguda a errors aleatoris (Kaaks 1995a). Els coeficients ϕ_Q i δ_Q representen el biaix constant (quan un individu tendeix a infra o supraestimar el seu consum de forma constant) i d'escalat proporcional (quan aquesta infra o supraestimació és proporcional al consum real) respectivament. Kaaks (1997) diu que ε_Q està correlacionat positivament amb T (o sigui, que els individus amb consum alt tindran errors positius més sovint i els de consum baix més errors negatius) cosa que provocarà sempre una infraestimació de β (aquesta infraestimació s'anomena biaix d'atenuació (Armstrong 1992)) i ve modulada per la correlació entre Q i T, ρ_{QT}. En canvi, δ_Q pot estar correlacionat positiva o negativament amb T. Si la covariància entre aquest error sistemàtic i T és positiva δ_Q serà major que 1 (sobreestimació de Q en els que consumeixen més). En aquest cas, la direcció del biaix que afecta β és la mateixa que la que provoca ε_Q (o sigui infraestimar β^*). En canvi, si la covariància fos negativa (sobreestimació de Q en els que consumeixen menys) δ_Q serà menor que 1, i causarà un biaix de sentit contrari al que provoca l'error aleatori, arribant, fins i tot a obtenir valors de $\lambda=1$ (es cancel·laria el biaix) amb presència de biaixos de diferents direccions deguts als errors aleatoris i sistemàtics (en aquest cas es pot demostrar que la covariància entre Q i l'error total de mesura e_Q és zero (errors berksonians (Berkson 1950) i que $\delta_Q = \rho_{QT}^2$)). A la pràctica, però, els valors de δ_Q s'espera que siguin al voltant de 1 i ρ_{QT}^2 al voltant de 0,5 , que provocarà que el biaix faci infraestimar β^* . Si multipliquem Q per un factor apropiat f, el paràmetre δ_Q canviarà a $f\delta_Q$ mentre que ρ_{QT}^2 no es veu afectat, de forma que podem aconseguir errors totals e_Q berksonians (o sigui, errors que no aporten biaix). Kaaks (1997) demostra que f ha de valer precisament λ (factor de calibratge) per fer desaparèixer el biaix. Cal notar, però, que la presència d'errors berksonians, si bé no dona biaix, sí que redueix la potència (més com més gran sigui ε_Q). Com que el biaix fa infraestimar β^* , al factor de calibratge λ també se l'anomena factor de desatenuació.

Si T, Q i ε_Q tenen una distribució normal es pot veure que (Kaaks 1995a)

$$X = E[T|Q] = \alpha' + \lambda Q \qquad (15)$$

amb $\alpha' = (\mu_T - \lambda\mu_Q)$ i $\lambda = \rho^2_{QT}/\delta_Q$. Aquesta expressió és exactament igual que dir que λ =cov(T,Q)/var(Q) (Kipnis 1999) [$= \rho_{QT}\frac{\sigma_T}{\sigma_Q}$ (Kipnis 2002)] que és la forma més habitual d'expressar la pendent d'una regressió lineal. Quan l'error dels qüestionaris *Q*, e_Q, està correlacionat amb *T*, λ pot ser negativa o major que 1, però en estudis nutricionals acostuma a estar entre 0 i 1 (Kaaks 1994b), ja que habitualment $\rho_{QT} \geq 0$ i $\sigma^2_T \leq \sigma^2_Q$ (Kipnis 2002). La distribució dels valors predits *X* serà aleshores normal, amb mitjana μ_T i variància $\rho^2_{QT}\,\sigma^2_T$ (Kaaks 1995a).

Substituint *X* per $\alpha' + \lambda Q$ a (13) obtenim una relació lineal entre *Q* i *D*, amb una pendent β esbiaixada per un factor λ [$\beta = \lambda\beta^*$]. El factor λ és la inversa del biaix d'escala proporcional δ_Q multiplicat per un factor d'atenuació ρ^2_{QT} que és el quadrat de la correlació entre el consum real i el mesurat. Notis també que la variància dels valors predits *X* és independent de δ_Q i que disminueix quan decreix la correlació entre *Q* i *T*. Com es veu només cal la pendent (λ) per corregir el RR, però també cal el terme independent (α') per comparar les ingestes corregides entre les diverses submostres (Stram 2000).

En estudis dietètics, habitualment no es pot mesurar directament *T*. Per estimar *X* fa falta un estudi addicional amb una mesura de referència *R* sense biaix (és a dir, que si repetíssim *R* molts cops, la seva mitjana representaria la ingesta habitual real de l'individu (Carroll 1996))

$$R = T + \varepsilon_R \qquad (16)$$

amb ε_R amb mitjana zero i independent dels errors de mesura del qüestionari general [$Cov(\varepsilon_R, \varepsilon_Q)=0$]. Si aquests errors són independents, no s'espera que ε_R causi biaix en l'estimació de λ (aquest vindria donat només per ε_Q), per tant les mesures de referència *R* no han de ser necessàriament precises i es poden basar en una única administració de la mesura de referència (Kaaks 1997). Notis que l'error total de *R* (e_R) coincideix amb l'error aleatori ε_R ja que *R* no té error sistemàtic. Sota aquestes assumpcions

E[R|Q]=E[T|Q] i els valors predits *X* es poden estimar a partir de la regressió de *R* sobre *Q* i la variància de *X* es pot obtenir calculant directament la variància dels valors predits per aquesta regressió (Carroll 1996). Kipnis (1999) afegeix que cov(R,Q)=cov(T,Q) sota les assumpcions del model de calibratge (és molt fàcil de veure si $Q=T+e_Q$ i $R=T+e_R$ amb $cov(e_R,T)=0$ i $cov(e_R,e_Q)=0$).

Assumint la normalitat conjunta de *T, Q* i ε_Q (com a (14)) ara podem estimar els valors predits *X* com (Kaaks 1995a)

$$X=E[R|Q]= \alpha' + \lambda Q \qquad (17)$$

(noteu l'analogia amb (15)); això vol dir essencialment un reescalament lineal dels qüestionaris basals *Q* (d'aquí ve el nom de calibratge). Una forma ràpida de veure per què les dades calibrades *X* tenen menor variància és que com que λ és habitualment inferior a 1, i $var(X)=\lambda \cdot var(Q)$ aleshores $var(X)<var(Q)$ (Willett 1998).

Spiegelman (2001) dóna un model general: $g[E(D|T)]= \alpha^* + \beta^* T$ amb g(.) una funció de *link* que linearitza la funció mitjana condicional. De fet, però, en el model no hi posem *T* sinó *X* (obtingut de regressar *R* en *Q*). Quan g(.)=E(D|X) tenim el model lineal, quan g(.)=logit[E(D|X)] tenim el model logístic. Quan $g[E(D|X)]=logI(t|X=0) + \beta X$, amb *I(t)* taxa d'incidència en el temps *t* i g[E(D|X)]=log[I(t|X)], aleshores tenim un model de regressió de riscos proporcionals de Cox. Els mateixos mètodes de correcció de Rosner (1990) per a la regressió logística són aplicables per a regressió lineal (Spiegelman 1997b).

6.1.3 ADAPTACIÓ DEL MODEL DE ROSNER A ESTUDIS DE COHORT

Plummer (1994) adapta el mètode de Rosner per treballar amb taxes d'incidència en comptes d'OR: si la taxa d'incidència de la malaltia (*a*) es relaciona amb el consum habitual *T* mitjançant un model log-lineal

$$log(a)= \alpha^* + \beta^* T \qquad (18)$$

(β^* és el logaritme del RR per una unitat de diferència en el consum real) assumint que la ingesta habitual de cada individu és constant, però que varia entre individus, i que dins de cada cohort els consums estan distribuïts normalment, a nivell cohort observem que (Prentice 1982)

$$\log(a_i) = \alpha^* + \beta^* \mu_i + (\beta^* \tau_i)^2/2 \qquad (19)$$

amb μ_i i τ_i mitjana i desviació estàndard del consum a la cohort *i*.

Si assumim que el darrer terme de (19) es pot ignorar, aleshores la relació entre la ingesta i la malaltia també és lineal a nivell cohort. Aquesta assumpció tindria sentit si $(\beta^* \tau_i)^2/2$ fos petit o variés poc entre les cohorts (es faria gairebé constant i seria absorbit per α^*). Sota aquestes circumstàncies, la informació sobre la relació malaltia-dieta a nivell ecològic és un reflex de la relació malaltia-dieta a nivell individual.

L'efecte $\hat{\beta}$ estimat en relacionar la taxa de malaltia amb l'exposició mesurada amb el mètode general *Q* (assumim que aquest valor sigui relativament constant al llarg del seguiment) ve esbiaixat per un factor λ, així que la relació real entre malaltia i l'exposició serà $\hat{\beta}\hat{\lambda}^{-1}$, havent obtingut $\hat{\lambda}$ de fer la regressió de *T* (o un mètode de referència vàlid *R*) respecte a *Q*. Un procediment alternatiu a haver de calcular $\hat{\beta}$, $\hat{\lambda}$ i haver de dividir-les és calcular directament l'efecte corregit $\hat{\beta}^*$ usant en el model (18) la ingesta calibrada *X* en comptes de *T* (Plummer 1994). Aquesta ingesta calibrada només és el resultat d'aplicar els valors predits per a tota la cohort usant la ingesta general *Q*, i aplicant els coeficients obtinguts de la regressió entre *R* i *Q* (vegeu que si es vol es pot incorporar l'efecte de variables de confusió).

Spiegelman (1997b) demostra que un model de riscos proporcionals de Cox es calibra exactament igual com fa Rosner (1990) amb la regressió logística. Cal que els residus de la regressió de *T* amb *Q* siguin normals, que l'error sigui no diferencial, que les pèrdues del seguiment siguin independents de l'exposició, que es tracti d'una malaltia rara, amb un RR moderat i un error de mesura baix. En canvi, Carroll (1990) diu que no cal assumir normalitat en els residus del calibratge.

Un exemple de com pot variar el RR (Schatzkin 2003): per λ=0,4 , un RR real de 2 s'observaria com $2^{0,4}$=1,27.

6.1.4 ENFOCAMENT INDIVIDUAL I ENFOCAMENT ECOLÒGIC: ESTUDIS MULTICÈNTRICS.

El principal objectiu de la l'estandardització d'una eina de mesura dietètica és minimitzar (les diferències en) l'error de mesura (Slimani 2000). En aquest apartat veurem com la precisió del risc relatiu es pot millorar tenint en compte també la variabilitat entre cohorts a més de la variabilitat intracohorts en un estudi multicèntric.

La limitada variabilitat interindividu dels patrons dietètics que s'observa en estudis d'àmbit nacional ha justificat la necessitat de fer grans estudis multicèntrics (Kaaks 1997). Aquest enfoc incrementa la variabilitat de la ingesta i la mida mostral de la cohort i fa possible contestar preguntes que no són resolubles en estudis monocèntrics o multicèntrics d'abast solament nacional. En canvi, els desavantatges d'estudiar una gran cohort multicèntrica són que cal usar un instrument econòmic per mesurar la dieta i problemes de comparabilitat de les dades entre centres (Kynast-Wolf 2002). Molts cops l'eina de mesura de la dieta s'adapta a les característiques de la població i els aliments de cada país. Així, malgrat l'existència d'un protocol comú per a l'estudi, hi haurà diferències per país en l'obtenció de les dades, que donen lloc a diferents estructures en l'error de mesura. Això pot impedir un ús directe de les mesures dietètiques en l'obtenció dels riscos del consum en relació amb una malaltia (Riboli 2000). Per superar aquests obstacles, el calibratge, basat en un qüestionari de mesura de més qualitat i aplicat uniformement a una submostra de tots els centres, permet una correcció dels riscos obtinguts mitjançant les variables originals (Kynast-Wolf 2002).

Quan treballem amb estudis multicèntrics, podem abordar la relació de l'exposició i la malaltia des d'una perspectiva ecològica o nivell cohort (relació entre el valor mig d'exposició a cada subcohort i la malaltia) o des d'un enfoc individual (relació entre l'exposició i la malaltia dintre de cada subcohort) (Plummer 1994). Si tots els individus d'una mateixa subcohort tinguessin la mateixa ingesta no caldria treballar a nivell individual. Anàlogament, si la mitjana de consum a cada subcohort fos la mateixa no

caldria treballar a nivell ecològic. Tots dos enfocs tenen mancances: el nivell ecològic no permet ajustar per variables de confusió (Greenland 1992) i s'ha demostrat la inconveniència de barrejar diferents cohorts sense tenir en compte, en fer l'anàlisi a nivell individual, que els individus procedeixen de diferents cohorts (Piantadosi 1998). Per exemple, és habitual trobar discrepàncies entre tots dos nivells. Podem trobar un factor associat a una malaltia a nivell individual, però degut a la falta de control sobre les variables de confusió no trobar-lo a nivell ecològic; o trobar la relació a nivell ecològic però no trobar-la a nivell individual pel fet que l'efecte ha estat atenuat per l'error de mesura. En un estudi multicèntric es poden obviar aquests problemes ajustant per les variables de confusió abans d'agregar les dades. A més, cal tenir en compte que el nivell d'error sistemàtic pot afectar de forma diferent a les diverses cohorts, més si tenim en compte que es poden haver usat diferents eines de mesura per mesurar la mateixa variable; i que l'error aleatori individual de mesura pot afectar també de forma diferent a cada cohort. El calibratge pot corregir aquests problemes.

Diferències en el biaix proporcional δ_{Qi} o en la correlació entre Q_i i T_i, ρ_{QTi}, entre les diverses cohorts donarà lloc a diferents graus de biaix en l'estimador del RR β_i per a cada cohort i (Kaaks 1995b). Com que utilitzem models de la família exponencial, el factor de biaix constant ϕ_Q no afecta a l'estimació de β (Kaaks 1995b). L'heterogeneïtat de l'estimador de β_i per cada cohort deguda a les diferències del biaix provocades pel diferent grau d'error en els qüestionaris Q_i així com la millora de la precisió de l'estimador global β pot ser corregida per l'ús de la mesura de referència R (via calibratge) (Kaaks 1994a).

El calibratge millorarà la comparabilitat dels diferents riscos relatius entre cohorts només si cada factor de calibratge λ_i s'estima amb prou precisió (Kaaks 1995b). Kaaks (1995b) demostra que el calibratge és més eficient incrementant el nombre d'individus en l'estudi de calibratge o incrementant la correlació entre R i Q.

Resumint, el calibratge permet corregir simultàniament les diferències entre cohorts quant a biaix d'atenuació (pels errors aleatoris en els qüestionaris basals) i els biaixos proporcionals d'escala δ_Q deguts a les correlacions intracohort entre els errors de mesura i els valors reals d'ingesta. La quantificació d'aquests guanys en precisió degut a

reduir l'heterogeneïtat intercohort dependrà del grau d'heterogeneïtat observat i si es pot assumir o no que els riscos relatius (reals) siguin idèntics entre les cohorts (Kaaks 1995b) (indicaria un model d'efectes fixos o aleatoris respectivament (Spector 1991)).

Kaaks (1994a) dóna una bona explicació de com estimar el RR a nivell ecològic i individual i com combinar-los. Un dels handicaps amb que s'han trobat els estudis sobre dieta i càncer és una exposició massa homogènia, que porta a una disminució de la potència estadística (Wynder 1987). Tenint en compte que la majoria de RR's entre els quintils superior i inferior d'ingesta no deuen ser majors de 4 per a la majoria d'aliments i nutrients i tipus de càncer, la presència d'errors de mesura encara faria més difícil detectar aquestes associacions. A més de l'increment de la mostra o de la precisió de la mesura podem, doncs, incrementar l'heterogeneïtat de l'exposició per augmentar-ne la potència.

Kaaks (1994a) considera un estudi multicèntric, compost per *j* cohorts. Assumint que dins de cada cohort hi ha una relació lineal entre el logaritme de la taxa d'incidència de la malaltia *D* i una exposició real subjacent *T*

$$D_i = \overline{D}_i + \beta_i^*(T - \overline{T}_i) \qquad (20)$$

(per simplicitat ignorem les possibles variables de confusió), les pendents d'aquestes relacions log-lineals, β_i^*, són exactament les mateixes que el logaritme del RR d'emmalaltir per una unitat de diferència en l'exposició a dins de cada cohort. Podem estimar β (treiem l'asterisc per indicar que es basa en un qüestionari *Q* i no en la ingesta real *T*), amb la variància de β_i donada per (Truett 1967):

$$Var(\hat{\beta}_i) = \frac{1}{c_i \hat{Var}_i(Q)} \qquad (21)$$

on c_i és el nombre de casos a la cohort *i* i $\hat{Var}_i(Q)$ és la variància del qüestionari de dieta de la cohort *i*. Podem combinar les diferents $\hat{\beta}_i$ obtingudes a cada cohort en una $\hat{\beta}_W$ que resumeixi les relacions observades entre l'exposició i el logaritme de la taxa

d'incidència dins de les cohorts, tenint en compte el pes de cada una de les $\hat{\beta}_i$, que vindrà donat per la inversa de la variància de cada una d'elles:

$$\hat{\beta}_W = \frac{\sum_{i=1}^{j} c_i \hat{Var}_i(Q)\hat{\beta}_i}{C\hat{Var}_W(Q)} \qquad (22)$$

amb *C* el nombre total de casos (sumant totes les cohorts) i

$$\hat{Var}_W(Q) = \frac{1}{C}\sum_{i=1}^{j} c_i \hat{Var}_i(Q) \qquad (23)$$

seria com la mitjana de les variàncies intracohorts de les mesures dels qüestionaris, ponderades pel nombre de casos que aporta cada cohort.

La variància de $\hat{\beta}_W$ es pot expressar com:

$$Var(\hat{\beta}_W) = \frac{1}{C\hat{Var}_W(Q)} \qquad (24)$$

És evident, comparant (21) i (24) que la variància de l'estimador pel conjunt de les cohorts serà menor que les variàncies individuals (ja que $C>c_i$). Per tant, si les diferents $\hat{\beta}_i$ són relativament similars (mateix efecte en totes les cohorts), combinant-les en $\hat{\beta}_W$ obtindrem un estimador més eficient per veure la relació entre l'exposició i la malaltia.

El problema és que aquesta homogeneïtat d'efectes entre les diverses cohorts, si fos real, es podria veure amenaçada per la presència de biaixos diferents en les cohorts deguts a variables de confusió, i a la presència de diferents factors que interactuïn amb l'exposició que modifiquin la probabilitat de patir la malaltia (modificació de l'efecte) (Kaaks 1994a). La inclusió de variables de confusió en el model resoldria el primer punt. La presència de modificadors de l'efecte indicaria que els riscos relatius associats a una certa exposició dependrien de característiques individuals distribuïdes diferentment entre les diverses cohorts, el que faria que les $\hat{\beta}_i$ no fossin combinables

(Greenland 1987). Una solució a aquest problema seria combinar les diferents $\hat{\beta}_i$ però en diversos estrats homogenis respecte al modificador de l'efecte.

La presència d'errors de mesura pot ésser una font addicional de variabilitat intercohorts, més encara si s'usen diferents tipus de qüestionaris (Kaaks 1994a). L'error de mesura, dins de cada cohort, es podria descriure com:

$$Q - \overline{Q}_i = \delta_i (T - \overline{T}_i + \varepsilon) \qquad (25)$$

A nivell ecològic (de grup) el biaix ve donat per $\overline{Q}_i - \overline{T}_i$. El coeficient δ indicaria el biaix proporcional (per exemple, tendència dels individus amb gran consum a infrareportar el consum). ε seria l'error aleatori, independent amb mitjana 0 i variància σ_ε^2.

Per millorar la comparabilitat entre cohorts s'haurien de corregir els riscos relatius afectats pel biaix d'error de mesura (Kaaks 1994a). Aquesta correcció (calibratge) ve donada per:

$$\hat{\beta}^* = \hat{\beta} / \hat{\lambda} \qquad (26)$$

(λ s'estima en una submostra en que es mesura *R* (en comptes de *T*), valor estimat no esbiaixat de *T* [R=T+ε_R], amb ε_R i ε independents, fent la regressió entre *R* i *Q*). *X* serien els valors predits (calibrats) a partir d'aquesta regressió.

La variància del nou estimador corregit ve donada per (Rosner 1989):

$$Var(\hat{\beta}_i^*) = \frac{1}{\hat{\lambda}_i^2} Var(\hat{\beta}_i) + \frac{\hat{\beta}_i^{*2}}{\hat{\lambda}_i^4} Var(\hat{\lambda}_i) \qquad (27).$$

Si la mostra de calibratge és prou grossa podem assumir que el segon sumand de (27) és gairebé 0, amb la qual cosa usant (21) en (27) podem escriure que:

$$Var(\hat{\beta}_i^*) = \frac{1}{\hat{\lambda}_i^2} Var(\hat{\beta}_i) = \frac{1}{c_i \hat{Var}_i(X)} \quad (28)$$

amb (Kaaks 1994a)

$$Var(X) = \rho_{QT}^2 Var(T) \quad (29).$$

El calibratge també millora l'eficiència de l'estimador intercohorts, com es pot veure substituint $Var_i(Q)$ per $Var_i(X)$ en (22):

$$\hat{\beta}_W = \frac{\sum_{i=1}^{j} c_i \hat{Var}_i(X) \hat{\beta}_i}{C\hat{Var}_W(X)} \quad (30)$$

de forma que les cohorts amb menys variabilitat en la ingesta real (*var(T)* petita) o amb correlació entre *Q* i *T* baixa tindran menys pes en la construcció de $\hat{\beta}_W$.

L'estimador $\hat{\beta}_W$ es basa únicament en comparacions entre individus pertanyents a una mateixa cohort (Kaaks 1994a). Es podria obtenir un estimador a nivell ecològic $\hat{\beta}_B$ fent la regressió del logaritme de les taxes d'incidència en cada cohort ($I_i=\log(c_i/n_i)$) respecte a l'exposició mitjana en cada cohort i ponderant per la precisió amb què s'estima cada valor (Prentice 1989). $\hat{\beta}_B$ podria resultar esbiaixat si els valors mitjos d'exposició $\overline{Q}_i$ són infra o sobreestimats de forma diferent en cada cohort. Una solució seria usar el valor real mitjà de consum de cada cohort a partir d'un qüestionari estandarditzat per a totes les cohorts, en una submostra representativa. La possible imprecisió del mètode es podria corregir usant mostres prou grosses. La precisió de cada punt $(\overline{T}_i, I_i)$ ve donada per la variància de I_i [$=1/c_i$], amb què obtenim

$$\hat{\beta}_B = \frac{Cov_B(T, I)}{\hat{Var}_B(T)} \quad (31).$$

Cal interpretar $Cov_B(T,I)$ com la covariància entre cohorts de les exposicions mitjanes T_i i les taxes d'incidència estimades I_i ponderada pel nombre de casos de cada cohort (Kaaks 1994a).

La variància de l'estimador intercohorts serà:

$$Var(\hat{\beta}_B) = \frac{1}{C\hat{Var}_B(T)} \qquad (32).$$

Si els estimadors intra $\hat{\beta}_W$ i intercohorts $\hat{\beta}_B$ del RR són prou semblants es poden combinar en un estimador global $\hat{\beta}_O$:

$$\hat{\beta}_o = \frac{\hat{Var}_W(X)}{\hat{Var}_W(X)+\hat{Var}_B(T)}\hat{\beta}_W + \frac{\hat{Var}_B(T)}{\hat{Var}_W(X)+\hat{Var}_B(T)}\hat{\beta}_B \qquad (33)$$

amb variància

$$Var(\hat{\beta}_O) = \frac{1}{C[\hat{Var}_B(T)+\hat{Var}_W(X)]} \qquad (34)$$

que demostra que la precisió del RR es pot millorar tenint en compte també la variabilitat entre cohorts a més de la variabilitat intracohorts.

6.1.5 JUSTIFICACIÓ PER USAR UNA SOLA MESURA DEL R24H COM A MÈTODE DE REFERÈNCIA

Com que el biaix d'atenuació depèn només de l'error aleatori de la variable predictora *Q*, no s'espera que els errors aleatoris de la mesura de referència *R* causin cap biaix en l'estimació de λ. Això justifica que es pugui prendre com a mètode de referència una sola mesura de la dieta mitjançant un mètode no esbiaixat, com el R24H, malgrat que pot ser una estimació molt poc fiable del consum habitual individual. Perquè el calibratge sigui prou precís cal un nombre bastant gran d'observacions, ja sigui

incrementant la mostra o el nombre de mesures per individu. Kaaks (1995a, 1995b) demostra que si disposem d'*N* observacions a partir d'*M* mesures en *Y* individus, la precisió s'optimitza agafant M=1 i Y=N. L'únic inconvenient d'aquesta estratègia és que no podem estimar separadament ρ^2_{QT}, δ_Q i σ^2_T. Així, no podem saber si el biaix en el RR observat és degut a errors de mesura aleatoris o al biaix d'escala proporcional, o si una variància petita del consum estimat és deguda a una baixa correlació entre les mesures del qüestionari *Q* i els valors reals *T*, o reflecteix una variància poblacional petita dels nivells reals de consum (Kaaks 1995a). O sigui, la pèrdua de potència deguda a l'error de mesura no pot ser estimada. Per estimar aquests paràmetres individualment calen almenys dues mesures (replicades o de diferent tipus, però almenys una ha de ser no esbiaixada i amb errors no correlacionats amb els errors de la resta d'instruments de mesura, habitualment un biomarcador) amb què comparar el qüestionari *Q* (Kaaks 1994b). Tot i així, quan estimem RR per diferències de consum no és necessari conèixer aquests estimadors individualment.

Recapitulant, es pot dir que el propòsit del calibratge és donar una estimació no esbiaixada del consum mig usant una mesura de referència (Plummer 1994). Per aconseguir aquesta estimació no esbiaixada cal:

1. Evitar biaix de selecció al seleccionar la mostra de calibratge.
2. Què el mètode de referència sigui no esbiaixat.

Com es veu, no cal que el mètode de referència sigui gaire fiable, ja que l'objectiu és caracteritzar bé la mitjana d'ingesta de la subcohort, no la d'un individu dins d'una subcohort. Si el mètode és poc fiable, però, l'error aleatori serà gran. Això es pot compensar usant una mostra prou gran (Plummer 1994).

Una alternativa a l'ús de mostres grans de calibratge seria replicar les mesures de referència, que permetria, a més, calcular la variabilitat individual en la mesura de referència. Jain (2003) diu que s'espera que la variabilitat entre múltiples R24H capturi la ingesta a llarg termini d'una persona, però que les entrevistes repetides poden fatigar la persona i fer que perdi interès. El mateix autor diu que si bé el procés de calibratge corregeix l'estimació puntual de l'efecte també fa aquest efecte més inestable (degut a

l'ampliació dels intervals de confiança). El fet de recollir moltes mesures de dieta d'un mateix individu podria motivar també un biaix de selecció, ja que caldria usar individus molt motivats per la dieta, que podria suposar que ρ_{QT} fos més forta en l'estudi de calibratge que en l'estudi principal, cosa que portaria a una sobreestimació del poder estadístic de l'estudi i una infraestimació dels biaixos d'atenuació (Kaaks 1995a). També podrien aparèixer, aleshores, tendències temporals sistemàtiques (Carroll 1997). El R24H és més barat, permet obtenir una alta taxa de resposta fins i tot en individus de baix nivell educacional (no cal escriure com en un diari dietètic) i l'entrevista dura poc (Morgan 1987). Com en qualsevol mesura de dieta, però, el R24H té errors i tendeix a infraestimar el consum d'aliments i nutrients (Buzzard 1998).

6.1.6 ALTERNATIVES AL CALIBRATGE LINEAL

Un mètode alternatiu al calibratge, d'ús molt més minoritari, ha estat el *convergent conditional score* de Huang (2001). Via simulació, la regressió de calibratge que hem usat té menor variància i una cobertura de probabilitat (percentatge dels cops en què l'interval al 95% de confiança de $\hat{\beta}^*$ inclou el veritable valor de β^*) similar a la del mètode de Huang. En canvi si *T* no és normal, el nostre mètode pot tenir més biaix. En qualsevol cas té un comportament molt millor que no fer cap correcció. Aquest nou mètode necessita una mostra gran per compensar les pèrdues d'eficiència.

Stürmer (2002) compara el mètode de calibratge lineal usat en aquest projecte amb un mètode semiparamètric (Robins 1995) per estudis cas-control. Mitjançant un estudi de simulació, el mètode de calibratge és tan bo (en termes de biaix i cobertura de probabilitat) com el de Robins i més eficient quan s'usa una mesura de referència (quan disposem del valor real *T* aleshores el mètode de Robins és millor en termes de biaix i cobertura). Stürmer (2002) també cita altres possibles alternatives al calibratge, com un enfocament amb pseudoversemblança (Carroll 1991), un enfocament *bootstrap* (Haukka 1995) i mètodes baiesians (Richardson 1993).

Hoffmann (2002) proposa un mètode de calibratge no lineal, però calen mesures de referència repetides per poder estimar la variància intraindividu que aquest mètode

necessita. Mitjançant el model no lineal s'assoleix la mateixa distribució (en termes de mitjana, desviació estàndard, simetria i curtosi). La resta d'assumpcions d'aquest mètode són les mateixes que les del model lineal. Hoffmann (2002) també compara el calibratge mitjançant regressió lineal amb models més directes: el model additiu [$X_{ij} = Q_{ij} + \overline{R}_j - \overline{Q}_j$] i multiplicatiu [$X_{ij} = Q_{ij} * \frac{\overline{R}_j}{\overline{Q}_j}$], que no milloren els resultats del lineal.

Freedman (2004) presenta el model de reconstrucció de moments, en el que el calibratge usant regressió lineal seria un cas particular. La diferència amb el model que hem usat és que els valors predits (calibrats) són els estimadors empírics baiesians amb variància conservada del valor real condicionat a la variable de resposta.

6.2 APLICACIÓ DEL MODEL DE CALIBRATGE A L'ESTUDI EPIC

El model de calibratge que aplicarem en aquest projecte serà un model de regressió lineal amb efectes fixos. La variable dependent serà el consum de carn mesurat amb el R24H. Com a variable explicativa hi haurà la interacció del país amb el consum de carn mesurat en el QFA/HD, de forma que obtindrem coeficients de calibratge diferents per a cada país, i les variables d'ajust: centre, edat al reclutament , alçada, pes i estació de l'any en què es mesura QFA/HD. D'aquesta forma es tenen en compte els factors que pensem que poden influir en la qualitat del QFA/HD i la relació amb el R24H. Els models s'executaran separats per sexe. El fet d'obtenir coeficients de calibratge per país permet tenir en compte l'especificitat geogràfica en termes de qualitat de mesura de la dieta i l'amplitut de la ingesta. No es calculen a nivell de centre en pro d'obtenir coeficients més estables a més d'assumir una previsible homogeneïtat entre centres de cada país, almenys quant a la qualitat de les mesures. Per assegurar que tots els dies de l'any hi són igualment representats, es ponderarà per la combinació estació astronòmica (primavera-estiu-tardor-hivern)–dia de la setmana (dilluns/dijous– divendres/diumenge).

Dels models de calibratge s'exclouran aquells que en el QFA/HD han reportat un consum habitual de carn de zero grams (no consumidors), ja que s'assumeix que l'error de mesura en els no consumidors és pràcticament inexistent. A aquests individus se'ls imputarà directament un valor de zero en la variable predita (calibrada).

A part del model original que acabem de presentar, es provaran uns altres models, amb transformacions de les variables d'ingesta (arrel quadrada i logaritme en base dos), exclusió dels consumidors per sobre del percentil 99, reagrupament dels consumidors de menys de 3 grams de carn per dia amb els no consumidors i ajust addicional per consum calòric.

Per cada model es calcularan les prediccions (variable calibrada), i per avaluar l'ajust, els residus estudentitzats i els coeficients de Cook i *leverage*.

6.3 MODEL DE MALALTIA

Per estudiar la relació entre el consum de carn i l'aparició de CG en la nostra cohort usarem un model de riscos proporcionals de Cox (1972). El model de Cox és flexible en el sentit que no cal especificar la distribució de la funció de supervivència.

Elecció de l'escala de temps: Quan l'edat és un factor determinant en l'aparició d'una malaltia (com ho és en el cas del càncer gàstric, ja que se sap que la incidència creix amb l'edat) és preferible usar-la com a eix temporal en els models de Cox (Korn 1997) , en comptes de posar el temps de seguiment directament i ajustar per edat. Per tant, l'edat serà l'eix de temps en els nostres models.

Estratificació de l'anàlisi: Dintre de la cohort poden haver-hi diversos grups amb funcions de risc basal diferents. Per tenir en compte això el model de Cox permet estratificar per aquests grups. En un estudi multicèntric és apropiat estratificar per centre, ja que pot variar la forma de recollir la informació de la variable resposta (com es detecten els casos de càncer gàstric), la mesura de l'exposició (dieta i variables d'ajust), diferent estructura poblacional (per edat, sexe, nivell educatiu,…), etc. El més correcte seria estratificar per cadascun dels centres participants, però alguns dels centres no han recollit encara cap cas. Això provoca que tot el centre no compti per a l'anàlisi. Aleshores, es perden individus a risc en el denominador en calcular les taxes. A més, si s'agafen estrats massa petits hi ha el perill que l'estimador de *hazard ratio* (HR) sigui molt inestable (Korn 1997). Si volem considerar aquests individus en el denominador al moment de calcular les taxes d'incidència haurem de fer agrupacions més grans (països). Per tenir en compte les diferències entre centres ajustarem per aquesta variable a part d'estratificar per país.

Es podria estratificar per altres variables. Habitualment també s'estratifica per l'edat. Provarem si hi ha canvis al estratificar per aquesta variable.

Variables tempo-dependents: En el nostre estudi no hi ha variables tempo-dependents (excepte les que per definició són temps, com edat o temps de seguiment) ja que només es disposa d'una mesura inicial. Tot i així es podria pensar que els individus que van ser diagnosticats en els primers anys de seguiment podrien haver canviat d'hàbits (entre els

quals la dieta) poc abans del reclutament, degut a malalties precursores o símptomes del seu CG, i per tant que la dieta habitual (o altres informacions) que reporten no fos l'exposició real que, potser, va provocar la malaltia. Això provoca un biaix en la informació que pot comportar resultats paradoxals (els individus malalts tenen una "millor" dieta que els sans, però en realitat han adoptat aquesta dieta "millor" perquè es trobaven malament (tot i no estar encara diagnosticats de CG)). Aquest biaix repercutirà en una infraestimació dels HR's. Per valorar l'efecte d'aquest biaix, realitzarem un anàlisi de sensibilitat, estudiant inicialment tots els casos i restringint, després l'anàlisi a aquells que han estat seguits almenys dos anys (bàsicament les persones que han estat seguides menys de dos anys són casos de CG o d'altres malalties greus o mortals). Es trien dos anys de forma arbitrària, prenent com a base l'opinió d'experts que creuen que dos anys és un temps suficient entre l'aparició de símptomes (que provocarien un canvi d'hàbits) i el diagnòstic de la malaltia.

Variables d'ajust: En els models s'inclouen variables d'ajust que habitualment s'utilitzen en estudis sanitaris de base poblacional, com el nivell educatiu, el consum de tabac, l'índex de massa corporal (IMC) i l'energia consumida. S'han triat aquestes variables d'acord amb la literatura (Nyrén 2002).

La variable gènere mereix una consideració apart. Se sap que els homes tenen una major incidència de CG que les dones, i que el seu consum dietètic és diferent. Cal estudiar si es pot usar un model amb el gènere com a variable d'ajust (i la seva interacció amb la dieta) o si els efectes són tan diferents entre homes i dones que cal fer servir dos models separats. Per poc que es pugui es mantindran els homes i dones en el mateix model a causa del nombre limitat de casos disponibles.

Cal recordar que al moment de calibrar s'han exclòs els no consumidors (els que tenen un valor 0 en el qüestionari QFA/HD) i se'ls ha assignat directament un valor zero en la variable calibrada. Cal controlar aquest fet per mitjà d'una variable indicadora

Correcció de la variància: També cal notar que al fer el model de Cox, no estem tenint en compte que la variable calibrada prové d'un model de regressió. Així, el HR estimat té una variància infraestimada. Per corregir aquesta variància es fa un procediment *bootstrap*. *Z* cops es fa el mostratge dels participants que disposen de dades del

qüestionari basal i del de referència, amb repetició (això vol dir que alguns individus poden estar repetits i uns altres desaparèixer en la mostra de calibratge de cada iteració). Això fa que obtinguem *Z* variables calibrades diferents. Estimem un model de Cox *Z* cops, cadascun amb una de les diferents variables calibrades obtingudes, i per a tota la cohort. Això dóna lloc a una col·lecció de *Z* HR's. Aleshores, podem estimar l'error estàndard (SE) corregit així (Rosner 2001):

$$SE_{corregit}(\hat{\beta}^*) = \sqrt{(\sum_{b=1}^{Z} \text{var}(\hat{\beta}_b^*)/Z) + \frac{1}{Z-1}\sum_{b=1}^{Z}(\hat{\beta}_b^* - \bar{\hat{\beta}}^*)^2} \qquad (35).$$

Com es veu, l'SE corregit es compon de dues parts: la variació del paràmetre en si (primer sumand dins de l'arrel, la mitjana de les variàncies de beta obtingudes en cada iteració) i la variació del paràmetre deguda al *bootstrap* (segon sumand dins de l'arrel, la variància de les *Z* estimacions de beta).

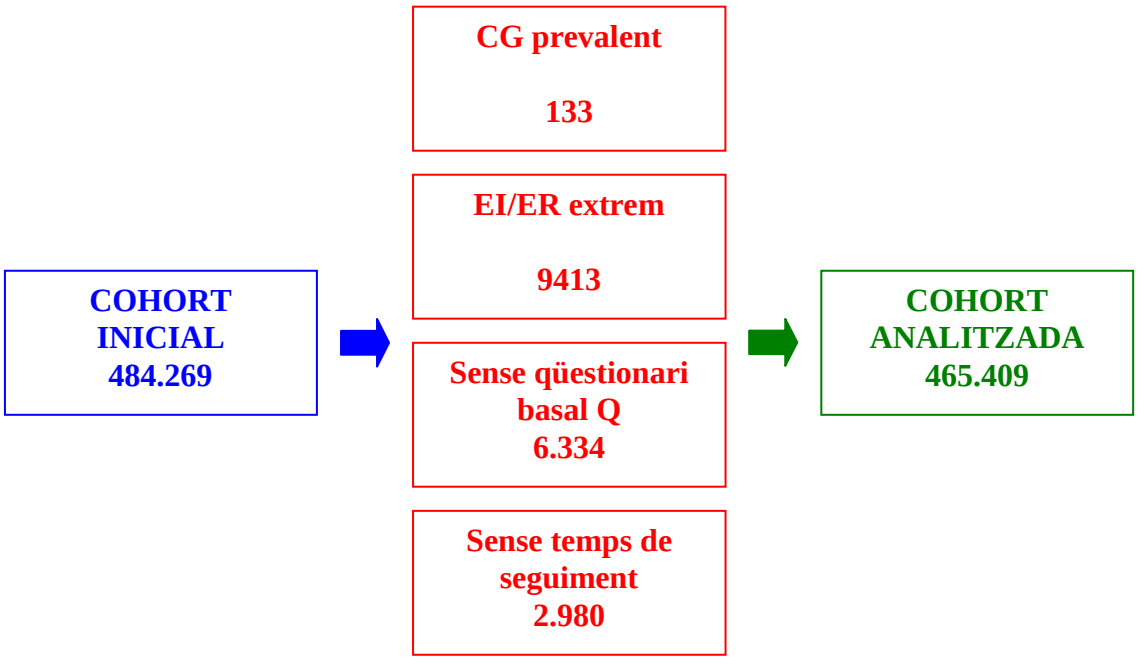
Bondat de l'ajust: Un cop hem ajustat el model de Cox, cal comprovar-ne la bondat de l'ajust. El model assumeix riscos proporcionals. Els residus de Schoenfeld, específics per a cada variable predictora, poden ajudar a detectar variables que no compleixin aquest supòsit. Si fem un diagrama de dispersió d'aquests residus (només es calculen per als casos, ja que té a veure amb l'aparició d'aquests) envers el temps de seguiment esperem no trobar-hi cap tendència (Therneau 2001). Els residus de martingala permeten veure si el model ajusta bé, tot i que no són pocs els autors que tenen reserves al respecte (Therneau 2001). En un bon model s'espera que els residus de martingala estiguin al voltant de zero. Una forma alternativa de veure l'ajust global del model és mitjançant els residus de Cox-Snell. Aquests residus han de tenir una distribució exponencial si el model de Cox ajusta bé les dades. S'espera que en una gràfica en què es creuin els residus de Cox-Snell amb el risc acumulat calculat usant els residus de Cox-Snell com a eix de temps i l'event original (CG) com a variable de censura, s'obtingui una recta de pendent 1 (Stata 2003). La detecció dels punts influents es realitzarà mitjançant l'estadístic LD (desplaçament de la versemblança al eliminar un punt) (SAS 2001).

7. RESULTATS

7.1 DESCRIPCIÓ DE LA MOSTRA

Els participants noruecs no van ser inclosos a l'estudi ja que la seva cohort va ser reclutada bastant més tard que la resta i per tant tenia un seguiment molt curt. Dels 484.269 individus restants se'n van excloure 18.860 per diverses causes (figura 3). Els que tenien una raó energia consumida/energia requerida (EI/ER) per sobre del percentil 99 o per sota del percentil 1 també van ser-ne exclosos. L'energia consumida es calcula a partir del qüestionari de dieta basal i l'energia requerida a partir de dades antropomètriques de l'individu. Teòricament, d'aquesta forma eliminem aquells individus que reporten un consum molt per sobre o molt per sota de les seves necessitats. A la taula 3 es poden veure com varien alguns valors segons si incloem o no el dos per cent extrem d'EI/ER. Bàsicament es pot veure com les mitjanes gairebé no varien, però que els màxims es redueixen i alguns valors impossibles (com energia zero) desapareixen.

Figura 3. Esquema de les exclusions en la cohort EPIC.



Taula 3. Canvis en les variables consum de carn i energia si incloem o excloem els individus amb un EI/ER inferior al percentil 1 o superior al percentil 99.

		Incloent EI/ER extrems (n=149.154+325.668)				Excloent EI/ER extrems (n=146.209+319.200)			
		Mitjana	SD	Mínim	Màxim	Mitjana	SD	Mínim	Màxim
Homes	Carn (g/dia)	127,3	75,7	0	3432	126,6	71,6	0	1196
	Energia (Kcal/dia)	2479,8	768,2	0	32909	2465,8	688,9	758	6260
Dones	Carn (g/dia)	89,8	56,4	0	3154	89,5	53,9	0	829
	Energia (Kcal/dia)	2002,2	608,7	0	27206	1991,5	547,4	614	5301

Finalment la mostra efectiva per a l'anàlisi és de 465.409 individus pertanyents a 9 països (taula 4), amb 34.456 individus que participen també en l'estudi de calibratge, que suposa més d'un 7% de la mostra general.

Taula 4. Distribució per països dels participants a l'EPIC i a l'estudi de calibratge.

	Estudi general			Estudi de calibratge			
	Homes	Dones	Total	Homes	Dones	Total	%
França	0	73025	73025	0	4665	4665	6%
Itàlia	14302	31389	45691	1418	2451	3869	8%
Espanya	15320	25290	40610	1743	1407	3150	8%
Regne Unit	24045	56659	80704	509	772	1281	2%
Holanda	10041	29041	39082	1250	3227	4477	11%
Grècia	10505	15170	25675	1395	1404	2799	11%
Alemanya	22346	29629	51975	2229	2114	4343	8%
Suècia	23014	29720	52734	2745	3264	6009	11%
Dinamarca	26636	29277	55913	1900	1963	3863	7%
TOTAL	146209	319200	465409	13189	21267	34456	7%

S'han detectat 270 casos de CG entre els 465.409 individus inclosos finalment a l'estudi. A la taula 5 es pot veure el nombre de casos per centre i gènere.

A la taula 6 es poden veure algunes característiques dels casos i els individus a risc. Cal dir que cap de les variables explicatives usades té valors mancants, excepte l'hàbit tabàquic, no disponible per 4.446 individus a risc i 2 casos. Per tant, per realitzar el calibratge disposarem de 34.456 individus i per als models de malaltia disposarem de 460.693 individus a risc i 268 casos.

Taula 5. Distribució per centres dels casos de CG.

	Homes	Dones
Nordest de França	-	4
Nordoest de França	-	2
Sud de França	-	2
Costa sud de França	-	1
Florència	7	11
Varese	3	15
Ragusa	3	2
Torí	6	3
Nàpols	-	0
Astúries	4	1
Granada	0	7
Múrcia	0	1
Navarra	12	2
Sant Sebastià	1	0
Cambridge	23	6
Oxford (vegetarians)	5	5
Oxford (població general)	0	1
Bilthoven	3	0
Utrecht	-	14
Grècia	8	7
Heidelberg	7	7
Potsdam	18	6
Malmo	27	16
Umea	10	7
Aarhus	2	2
Copenague	6	3
TOTAL	145	125

Taula 6. Distribució de l'hàbit tabàquic, nivell d'estudis, edat al reclutament, IMC, consum de carn basal i ingesta energètica per a casos i individus a risc segons gènere.

	Homes				Dones			
	A risc		Casos		A risc		Casos	
	n	%	n	%	n	%	n	%
HÀBIT TABÀQUIC								
Mai fumador	48624	33,5%	35	24,5%	187183	59,3%	63	50,4%
Ex-fumador	53943	37,2%	60	42,0%	70411	22,3%	33	26,4%
Fumador	42624	29,4%	48	33,6%	57908	18,4%	29	23,2%
No disponible	873		2		3573			
TOTAL	145191	100,0%	143	100,0%	315502	100,0%	125	100,0%
NIVELL D'ESTUDIS								
Sense estudis	5815	4,0%	4	2,8%	14915	4,7%	9	7,2%
Primària	39218	26,8%	63	43,4%	68898	21,6%	45	36,0%
FP	35582	24,4%	33	22,8%	63797	20,0%	22	17,6%
Secundària	23070	15,8%	14	9,7%	81180	25,4%	27	21,6%
Universitat	38246	26,2%	28	19,3%	75887	23,8%	17	13,6%
No especificat	4133	2,8%	3	2,1%	14398	4,5%	5	4,0%
TOTAL	146064	100,0%	145	100,0%	319075	100,0%	125	100,0%
(*)	**Mitjana**	**SD**	**Mitjana**	**SD**	**Mitjana**	**SD**	**Mitjana**	**SD**
EDAT (anys)	52,37	10,19	60,26	7,76	51,34	10,20	58,97	8,68
IMC (Kg/m^2)	26,61	3,66	26,50	3,42	25,20	4,51	25,95	4,20
CARN (g/dia)	126,60	71,63	126,87	62,84	89,53	53,89	97,44	47,83
ENERGIA (Kcal/dia)	2465,84	688,94	2377,68	630,94	1991,56	547,39	1944,02	542,63

* El nombre d'individus per a les variables contínues coincideix amb els del total del nivell educacional.

7.2 L'ESTUDI DE CALIBRATGE

A l'estudi de calibratge es pondera pel dia de la setmana (agrupant dilluns-dijous i divendres-diumenge) i per l'estació de l'any en que es fa el R24H, per cada país i gènere. Els pesos varien entre 0,46 i 5,75 (menys d'un 10% dels ponderals valen menys de 0,5 o més de 2,0). Repetint l'anàlisi de calibratge sense pesos, els coeficients de desatenuació són prou semblants (no mostrat).

Com a mesura de precaució, no es calibraran aquelles variables en els centres en què les mitjanes del qüestionari general i del R24H tenen una raó inferior a 0,5 o superior a 2,0. Això indicaria que probablement els qüestionaris no mesuren el mateix (per exemple, degut a la inclusió d'un aliment força consumit en només un del qüestionaris). Podem veure quines són les mitjanes de consum de carn dels qüestionaris basals i de R24H per centre i gènere, així com la raó de mitjanes (taula 7). Aquesta raó varia entre el 0,62 dels homes d'Umea i el 1,68 de les vegetarianes d'Oxford. Així, el consum de carn té mitjanes prou similars entre els diferents qüestionaris, cosa que permet incloure tots els centres en el model de calibratge.

Si tenim en compte que les mitjanes del R24H ens donen la millor estimació de la ingesta d'un aliment a nivell grupal, a la taula 7 podem veure que el rang de consum de carn entre els centres és molt ampli. Sant Sebastià té el consum més elevat (242 i 127 grams/dia respectivament, per a homes i dones) mentre que Grècia té el consum més baix (77 i 46 grams/dia respectivament), si no considerem la cohort vegetariana d'Oxford.

Un cop efectuat el calibratge podem comparar els valors de la variable original (QFA/HD), de la variable de referència (R24H) i de la variable predita que s'usarà en el model de malaltia (variable calibrada) (taula 8). Ràpidament podem veure que un dels efectes del calibratge és un "encongiment" de les dades. Com calia esperar, la variable del R24H és la que té més variància, pel fet que es basa en el consum d'un dia. Els extrems de la variable calibrada també són més suaus que els de les variables original i de referència. Cal destacar la presència de valors negatius en la variable calibrada, deguda a l'efecte de les covariables. Recordem que als no consumidors (aquells que

reporten 0 grams a l'estudi basal se'ls assigna un zero directament, però no així als molt baixos consumidors. Aleshores l'efecte, fins i tot lleu, d'una covariable pot portar a prediccions negatives). Cal dir que, en aquest cas, els valors negatius només es donen a la cohort "vegetariana" d'Oxford, on els consums de carn són molt baixos.

Taula 7. Consum mitjà (g/dia) de carn usant el QFA/HD, el R24H i la raó entre totes dues mesures, segons centre i gènere.

	Homes			Dones		
	QFA/HD	R24H	Raó de mitjanes	QFA/HD	R24H	Raó de mitjanes
Nordest de França	-	-	-	114,47	107,00	1,07
Nordoest de França	-	-	-	102,04	107,54	0,95
Sud de França	-	-	-	100,45	101,35	0,99
Costa sud de França	-	-	-	97,12	99,73	0,97
Florència	129,41	130,57	0,99	102,68	95,09	1,08
Varese	127,46	154,17	0,83	98,50	89,10	1,11
Ragusa	103,58	122,12	0,85	78,03	79,53	0,98
Torí	117,82	125,73	0,94	94,01	91,67	1,03
Nàpols	-	-	-	88,49	65,48	1,35
Astúries	153,09	181,68	0,84	106,44	112,26	0,95
Granada	129,15	131,88	0,98	77,34	71,72	1,08
Múrcia	136,57	134,49	1,02	101,67	85,37	1,19
Navarra	177,03	173,59	1,02	126,05	108,94	1,16
Sant Sebastià	175,52	242,41	0,72	109,64	126,93	0,86
Cambridge	96,66	103,40	0,93	96,93	76,65	1,26
Oxford (vegetarians)	17,87	19,21	0,93	27,87	16,63	1,68
Oxford (població general)	97,20	107,83	0,90	97,31	65,87	1,48
Bilthoven	139,44	160,37	0,87	94,97	93,87	1,01
Utrecht	-	-	-	88,98	90,49	0,98
Grècia	84,34	76,89	1,10	68,98	46,01	1,50
Heidelberg	116,50	156,87	0,74	75,25	89,51	0,84
Potsdam	137,12	153,16	0,90	88,32	83,47	1,06
Malmo	127,98	134,17	0,95	89,71	91,95	0,98
Umea	82,46	133,75	0,62	60,45	88,05	0,69
Aarhus	178,37	138,21	1,29	111,42	86,91	1,28
Copenague	169,43	140,27	1,21	109,38	84,66	1,29

A la taula 9 es poden veure els coeficients de calibratge per a cada país amb els intervals de confiança. Cal recordar que cada centre, a més, té un terme independent propi. Per valors de $\lambda<0,2$ o $\lambda>1,0$ els models de calibratge poden ser massa inestables (relació massa dèbil entre les dues variables de mesura de la dieta). Tots els factors de calibratge per carn se situen entre 0,27 i 0,73. Són una mica més grans per als homes (mitjana no ponderada de tots els factors de calibratge=0,48) que per les dones (mitjana no ponderada=0,38, 0,39 excloent-ne França). Holanda i Alemanya són els països amb

factors de calibratge més alts, tant en homes com en dones (mitjana no ponderada de 0,59) mentre que Dinamarca en homes i Grècia en dones tenen els coeficients menors (0,27). El Regne Unit (mostra de calibratge més petita) i Grècia tenen els errors estàndard majors.

Taula 8. Consum de carn (g/dia) obtingut mitjançant el QFA/HD, R24H i valor predit (calibrat) segons gènere i país.

	QFA/HD					R24H					Dades calibrades				
HOMES	**n**	**Mitjana**	**SD**	**Mínim**	**Màxim**	**n**	**Mitjana**	**SD**	**Mínim**	**Màxim**	**n**	**Mitjana**	**SD**	**Mínim**	**Màxim**
Itàlia	14302	122,14	58,36	0	618	1418	132,63	110,65	0	663	14302	134,14	25,60	0	315
Espanya	15320	160,18	70,30	0	860	1743	183,95	147,21	0	947	15320	183,55	49,75	0	545
Regne Unit	24045	73,02	61,64	0	927	509	86,50	98,24	0	687	24045	74,88	46,79	-15	545
Holanda	10041	139,35	62,09	0	729	1250	160,37	125,69	0	1150	10041	161,48	47,14	0	609
Grècia	10505	97,90	45,53	0	442	1395	76,89	100,16	0	750	10505	86,06	28,41	0	279
Alemanya	22346	132,13	71,70	0	875	2229	154,84	121,59	0	1132	22346	162,19	42,72	0	600
Suècia	23014	110,61	58,61	0	651	2745	133,96	99,44	0	800	23014	140,32	23,61	0	367
Dinamarca	26636	173,71	66,36	0	1196	1900	139,66	95,34	0	791	26636	144,67	19,84	0	432
DONES															
França	73025	106,34	57,97	0	568	4665	104,42	85,44	0	813	73025	105,15	19,62	0	249
Itàlia	31389	96,38	44,99	0	477	2451	87,08	85,38	0	710	31389	89,04	23,19	0	262
Espanya	25290	107,04	51,05	0	693	1407	100,00	92,19	0	575	25290	102,56	26,20	0	274
Regne Unit	56659	56,65	56,77	0	685	772	57,28	70,24	0	555	56659	39,61	29,40	-12	276
Holanda	29041	92,45	47,26	0	435	3227	91,94	73,72	0	483	29041	93,11	27,93	0	303
Grècia	15170	73,76	33,68	0	378	1404	46,01	67,64	0	456	15170	50,77	11,88	0	139
Alemanya	29629	84,98	48,63	0	829	2114	86,54	81,51	0	578	29629	90,78	23,68	0	444
Suècia	29720	78,20	38,80	0	407	3264	90,05	72,36	0	478	29720	91,81	17,07	0	229
Dinamarca	29277	110,13	45,47	0	699	1963	85,23	69,59	0	436	29277	89,33	16,07	0	284

Taula 9. Coeficients de calibratge (λ), error estàndard (SE) i interval de confiança al 95% (IC95%), per país i gènere.

	Homes				Dones			
	λ	**SE**	**IC95%**		**λ**	**SE**	**IC95%**	
França	-	-	-	-	0,30	0,02	0,26	0,34
Itàlia	0,37	0,05	0,27	0,48	0,45	0,04	0,38	0,52
Espanya	0,44	0,04	0,37	0,52	0,33	0,04	0,24	0,41
Regne Unit	0,51	0,10	0,31	0,72	0,33	0,06	0,22	0,44
Holanda	0,73	0,05	0,63	0,83	0,57	0,03	0,51	0,62
Grècia	0,56	0,08	0,41	0,70	0,27	0,06	0,16	0,39
Alemanya	0,58	0,03	0,51	0,65	0,47	0,04	0,40	0,54
Suècia	0,39	0,04	0,31	0,47	0,39	0,04	0,31	0,47
Dinamarca	0,27	0,04	0,19	0,35	0,33	0,04	0,25	0,41

A les figures 4 i 5 per a homes i dones respectivament, es pot apreciar com canvia el percentatge d'individus dins de cada categoria de consum de carn en calibrar. Aquestes

categories es basen en els quartils específics per gènere de les variables de dieta originals *Q*, usant tota la cohort. Com es veu a les gràfiques, el calibratge fa que la majoria d'individus es desplacin a les categories centrals, de forma heterogènia entre països. Cal recordar que els punts de tall usats per la variable calibrada continuen essent els punts de tall dels quartils de la variable original *Q*. En alguns casos, però, s'observa un desplaçament dels individus cap a les categories de més consum en alguns centres (per exemple pels homes d'Espanya). Sembla que això és degut a que aquests països tenen centres amb termes independents bastant grans (el que provoca un desplaçament generalitzat dels individus cap a la dreta en calibrar).

Figura 4. Distribució del consum de carn per país en 4 categories, usant com a punts de tall els quartils específics per gènere calculats a nivell de tota la cohort a partir de la variable QFA/HD (Q1, Q2, Q3, Q4), segons si usem la variable del QFA/HD (Q) o la variable calibrada (X). Homes.

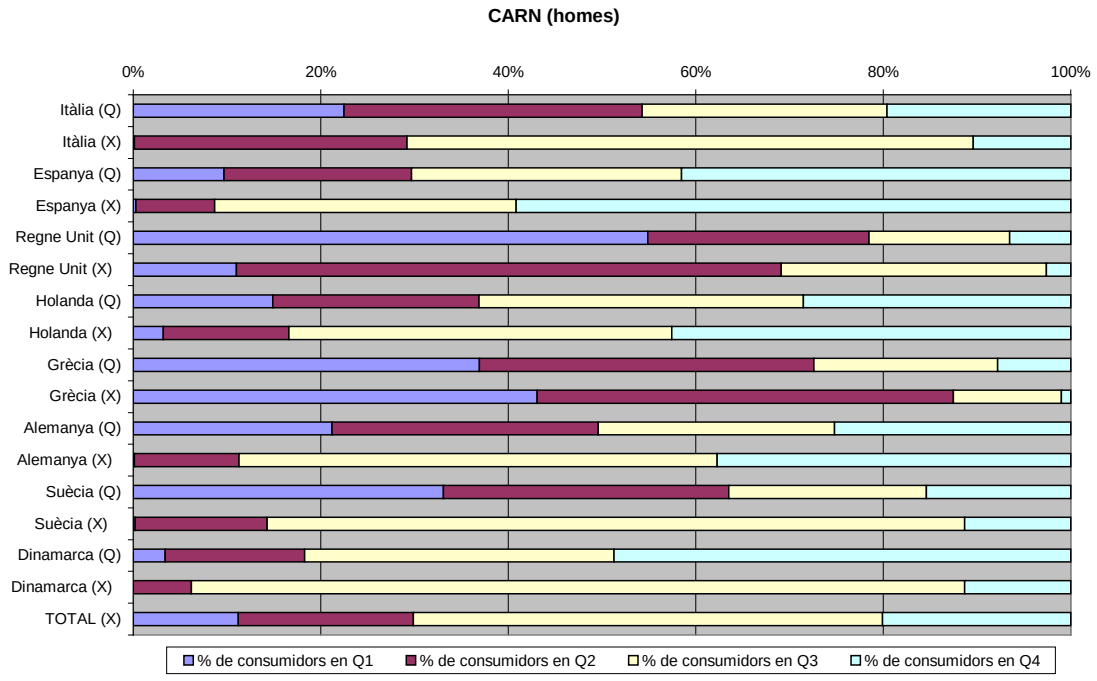
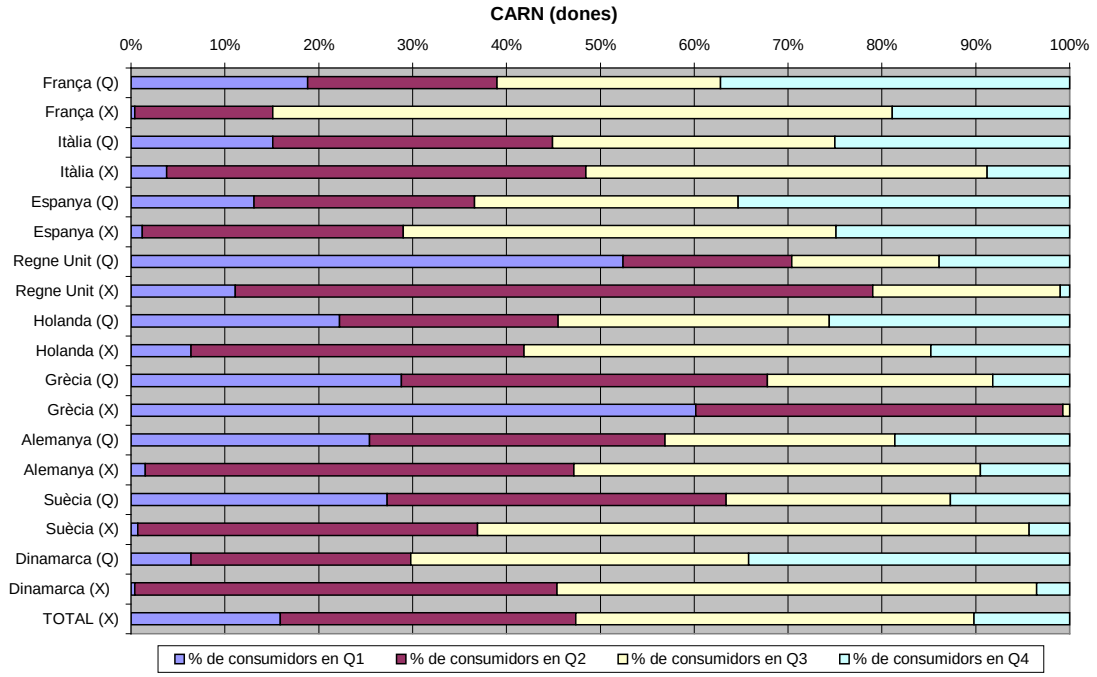


Figura 5. Distribució del consum de carn per país en 4 categories, usant com a punts de tall els quartils específics per gènere calculats a nivell de tota la cohort a partir de la variable QFA/HD (Q1, Q2, Q3, Q4), segons si usem la variable del QFA/HD (Q) o la variable calibrada (X). Dones.



A les figures 6 i 7 podem veure el mateix, però ara, per a la variable calibrada usant els quartils de la variable calibrada (calculats usant tota la cohort per a cada gènere). Com es pot veure, alguns centres canvien força de distribució respecte a la cohort després de calibrar, si bé els canvis no són tan espectaculars com quan usàvem els punts de tall de la variable original. Per exemple, els homes grecs tenien un consum baix de carn en el 37% de la mostra. Després de calibrar la proporció de baixos consumidors (respecte a tota la cohort) és del 84%.

Figura 6. Distribució del consum de carn per país en 4 categories, usant com a punts de tall els quartils específics per gènere calculats a nivell de tota la cohort a partir de la variable QFA/HD (Q1, Q2, Q3, Q4), si usem la variable del QFA/HD (Q) i a partir de la variable calibrada (X1, X2, X3, X4) si usem la variable calibrada (X). Homes.

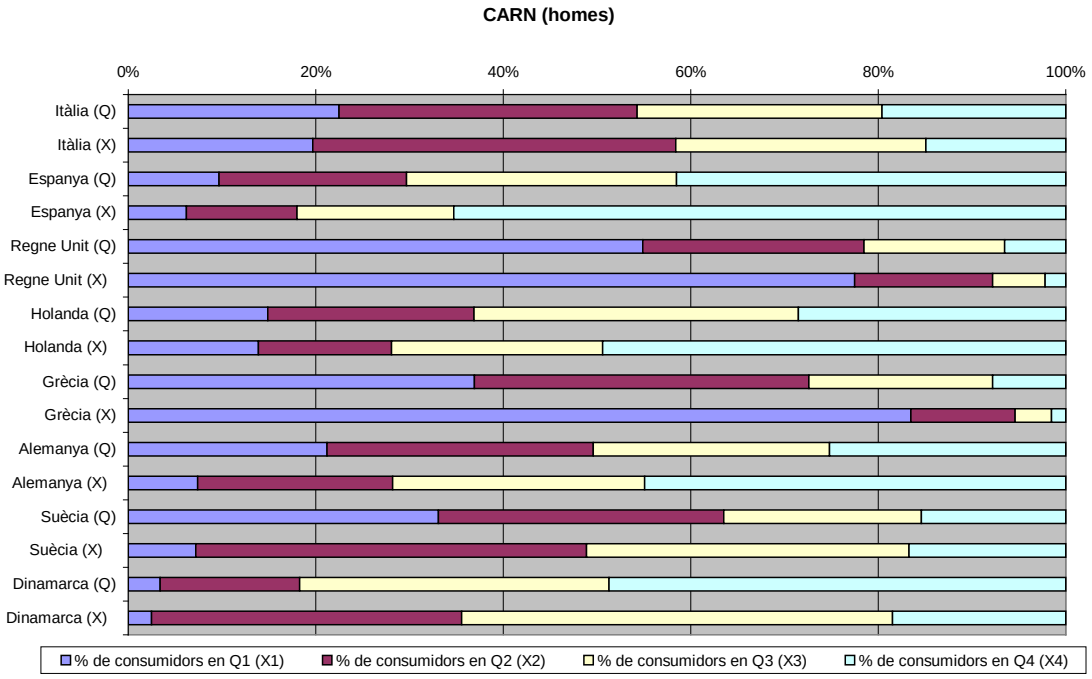
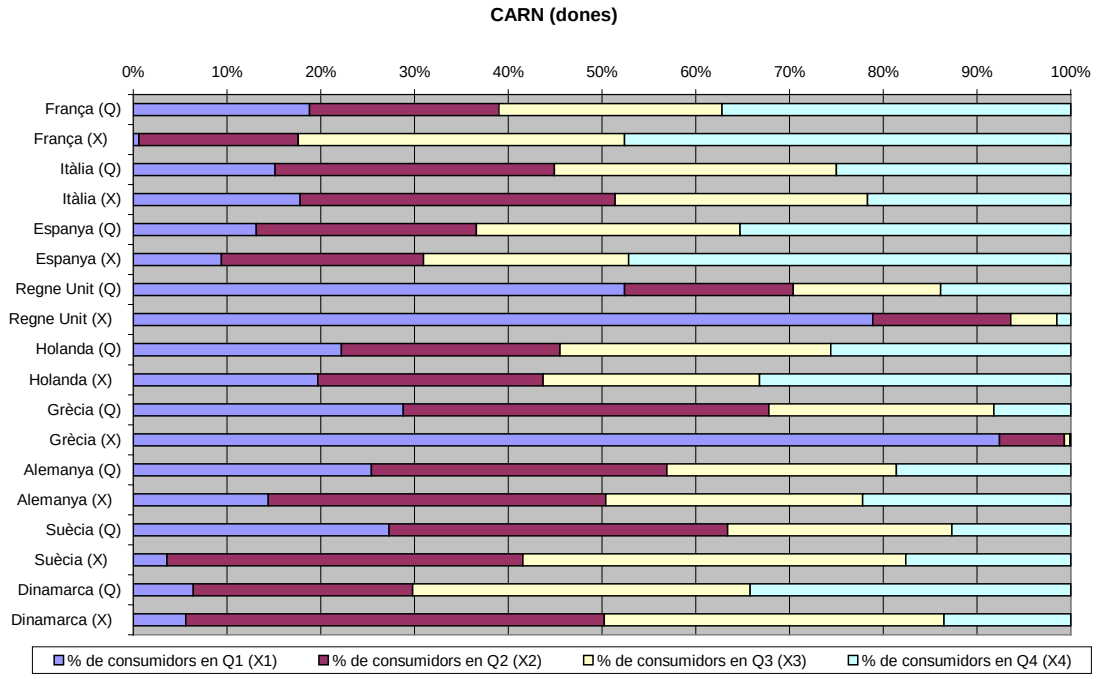


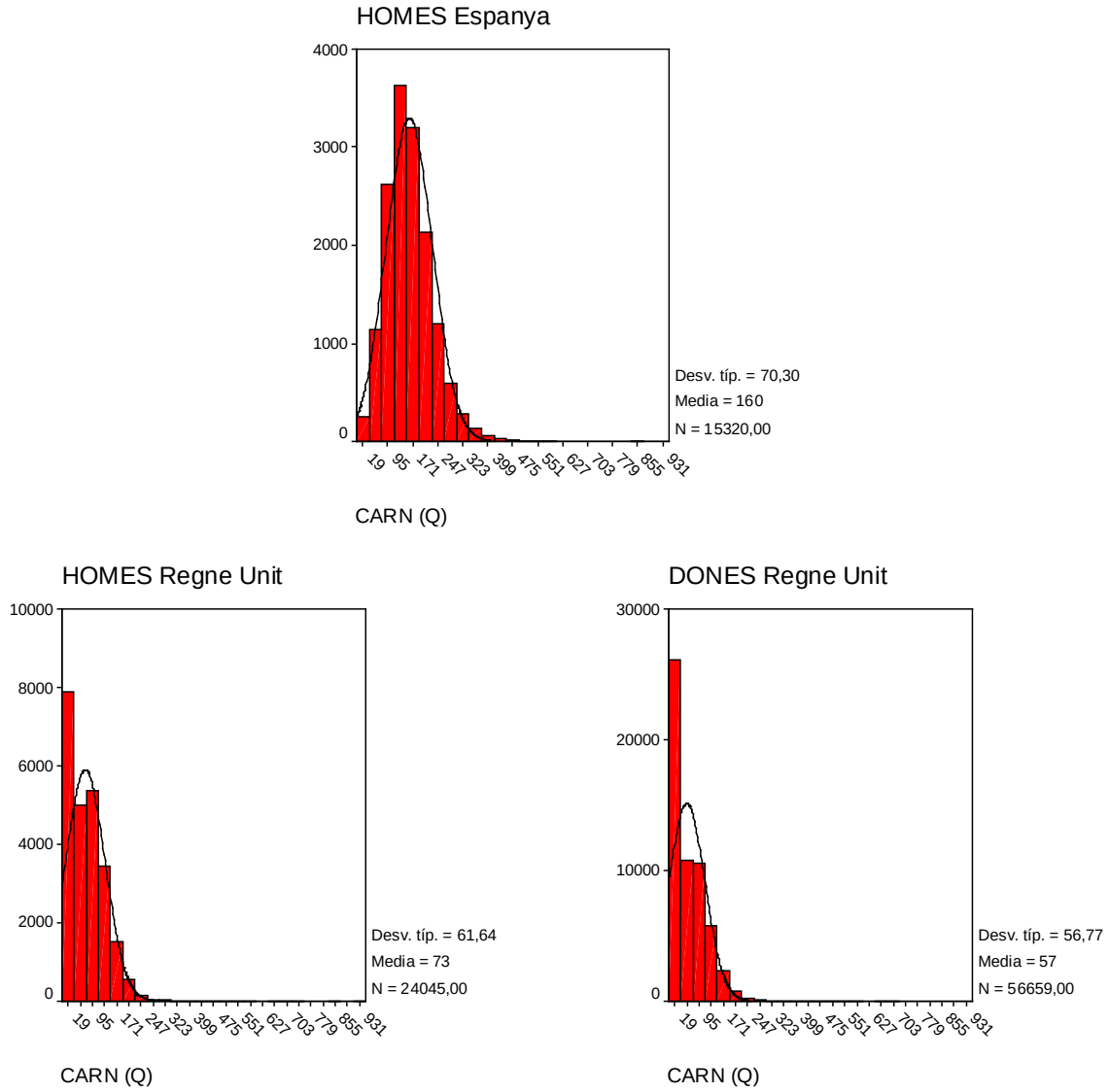
Figura 7. Distribució del consum de carn per país en 4 categories, usant com a punts de tall els quartils específics per gènere calculats a nivell de tota la cohort a partir de la variable QFA/HD (Q1, Q2, Q3, Q4), si usem la variable del QFA/HD (Q) i a partir de la variable calibrada (X1, X2, X3, X4) si usem la variable calibrada (X). Dones.



7.3 AJUST DEL MODEL DE CALIBRATGE

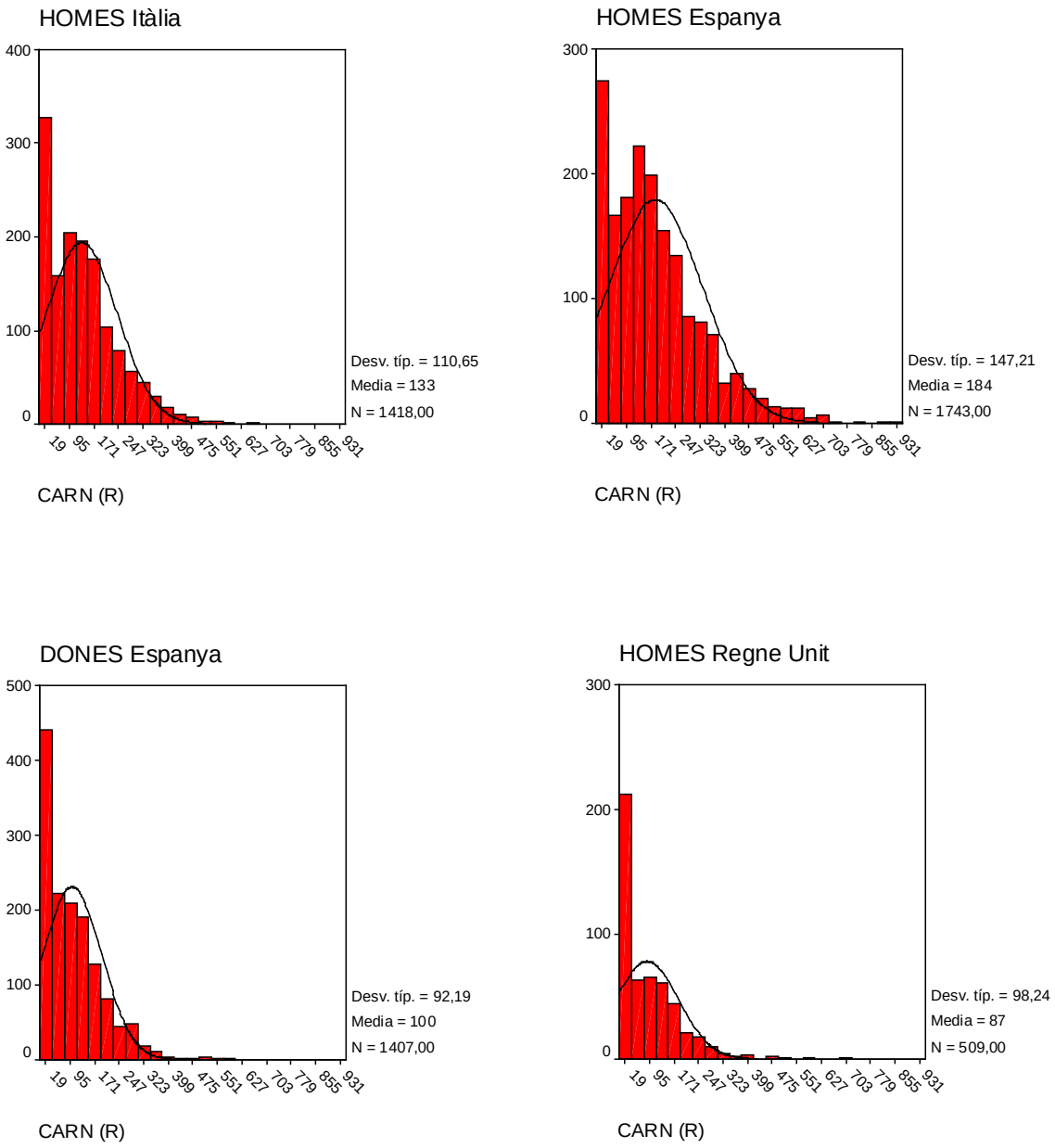
A la figura 8 podem veure els histogrames del consum habitual de carn *Q*. Per simplicitat només es mostren les gràfiques referides als homes espanyols, i a aquells centres, tan homes com dones, en què el comportament és prou diferent del dels homes espanyols. Com es veu, en la majoria de casos existeixen cues prou llargues per la dreta. Despreciant aquestes cues, la distribució és prou simètrica, excepte al Regne Unit.

Figura 8. Histogrames del consum de carn mesurat amb el QFA/HD per als homes d'Espanya i per a la cohort del Regne Unit.



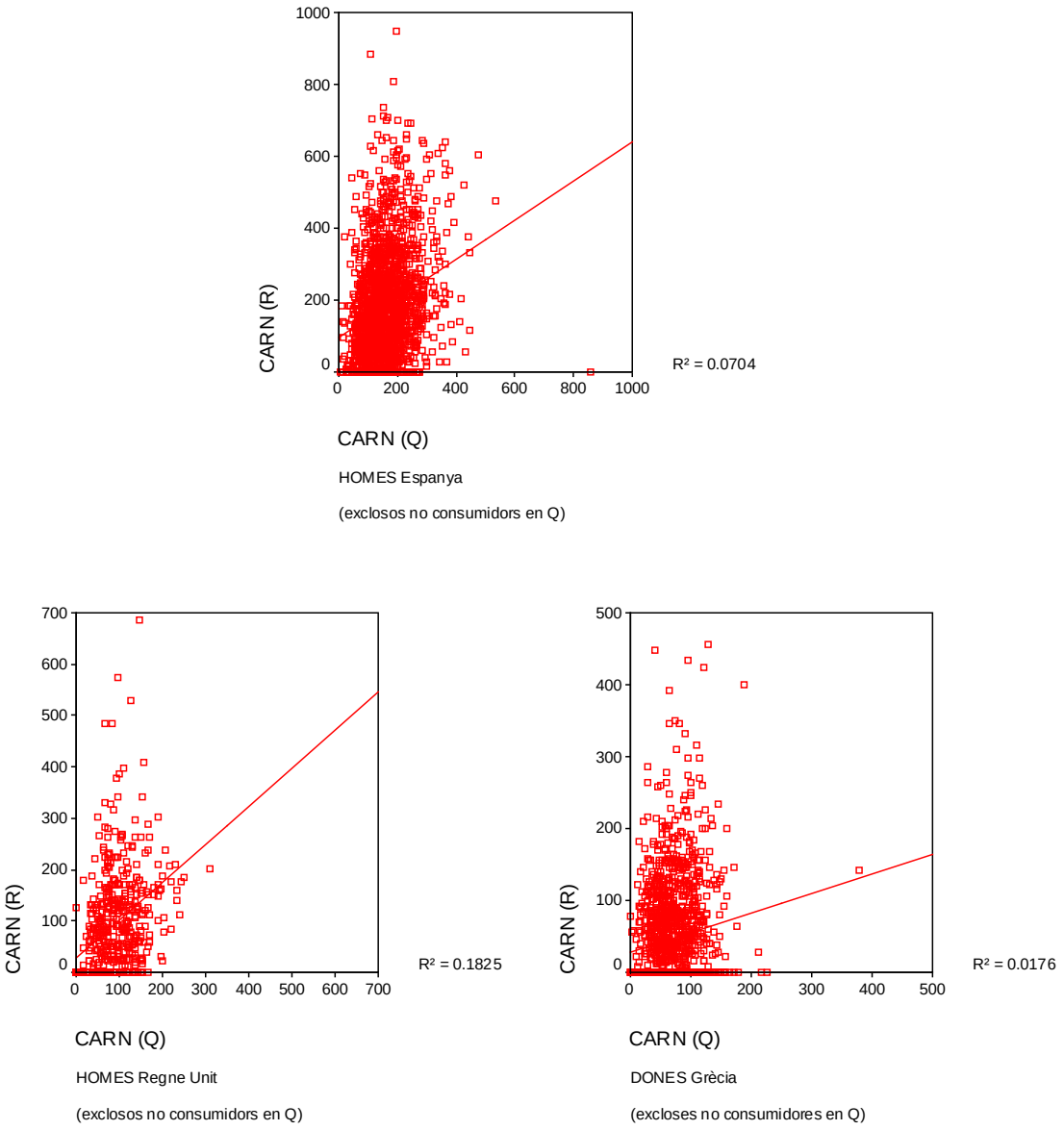
A continuació (figura 9) podem veure els histogrames corresponents al consum de carn mesurat amb el R24H. En aquest cas la distribució és clarament asimètrica, ja que acostuma a donar molt de pes als zeros (ja que es refereix al consum d'un sol dia). Tots els centres es comporten com Espanya aproximadament, excepte Itàlia (França té el mateix aspecte) i el Regne Unit (com Grècia). Les dones tenen una proporció més elevada de no consumidores.

Figura 9. Histogrames del consum de carn mesurat amb el R24H per als homes d'Itàlia, Espanya i Regne Unit i per a les dones d'Espanya.



Els gràfics de dispersió entre el consum de carn basal *Q* i el de referència *R* ens donen una idea de la relació entre les dues mesures. Es mostren els gràfics (figura 10) per als homes espanyols i els dos centres amb coeficients de determinació (r^2) extrems. En tots els centres la forma que defineix el núvol de punts és difícil d'identificar i l'ajust d'una recta dóna coeficients d'r^2 molt baixos, i varia entre 0,02 per a les dones de Grècia i 0,18 per als homes britànics. L'ajust de polinomis de segon i tercer grau no millora la situació. Cal recordar que els no consumidors habituals (60 individus) n'han estat exclosos, ja que se'ls assigna directament el valor 0. També s'observen valors molt alts de consum (possibles *outliers*).

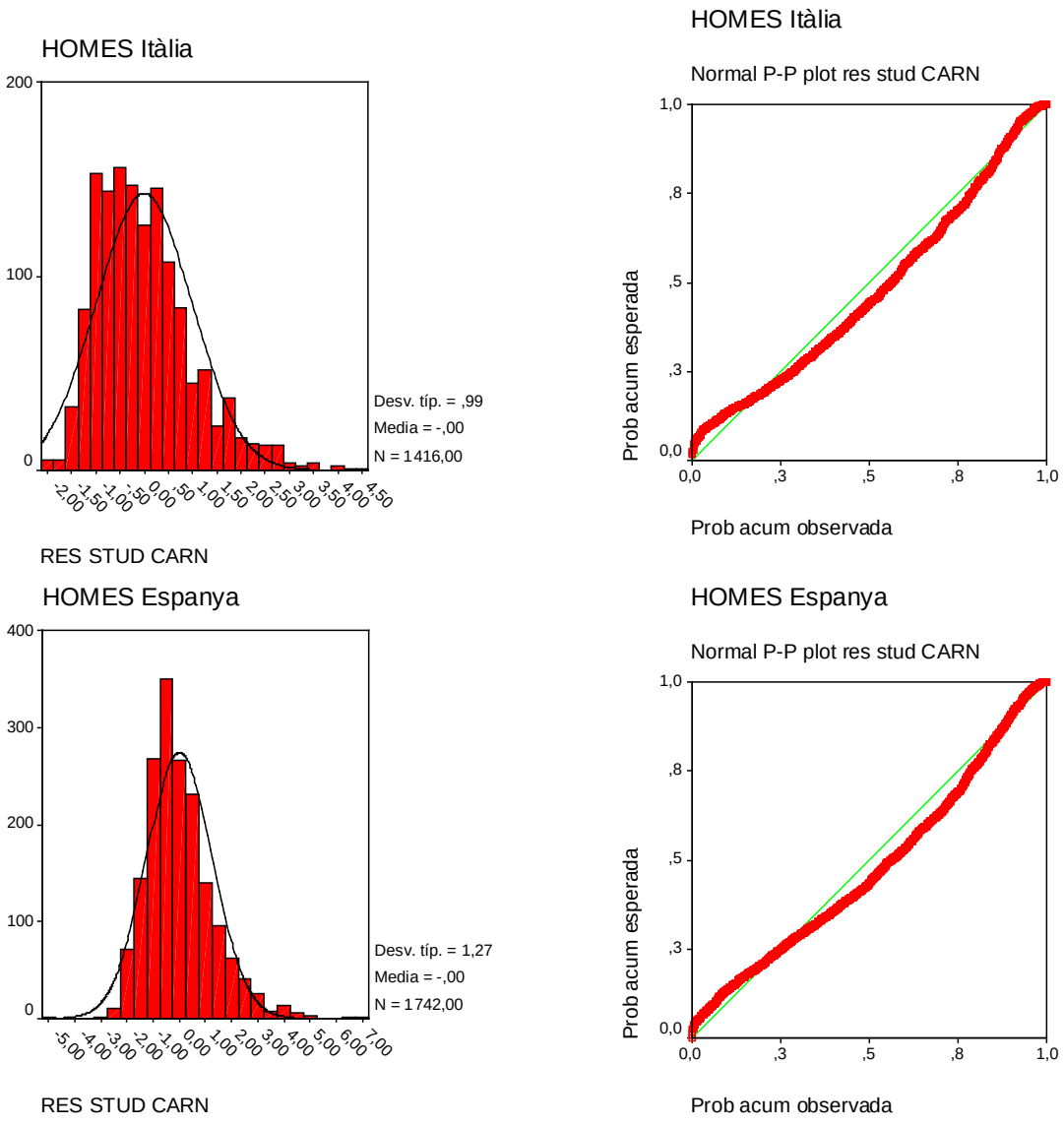
Figura 10. Diagrames de dispersió i coeficient determinació entre el consum de carn mesurat amb el R24H (R) i el QFA/HD (Q) per als homes d'Espanya i Regne Unit i per a les dones de Grècia.

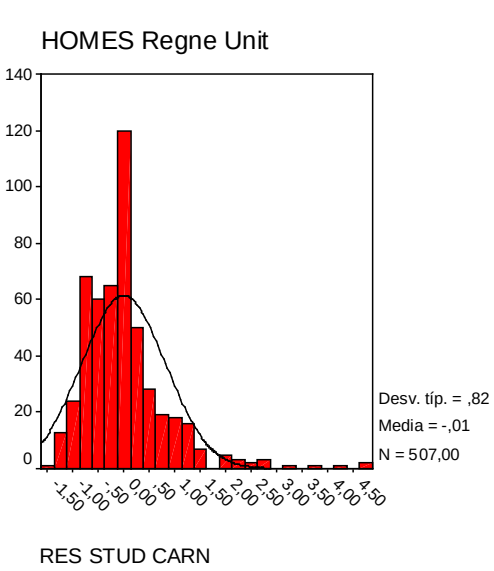


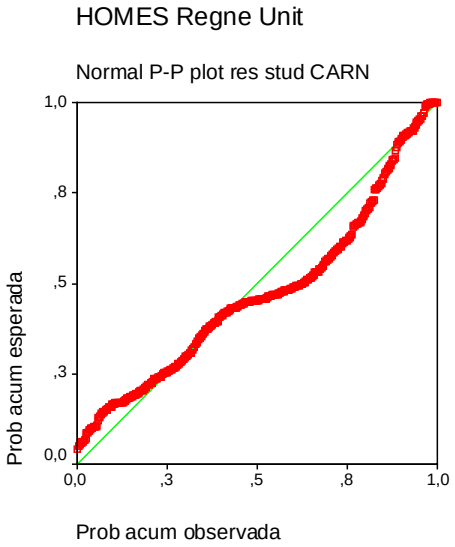
Una anàlisi gràfica dels residus hauria de permetre comprovar les hipòtesis de linealitat, normalitat, homocedasticitat i independència. S'usaran els residus estudentitzats, que són menys sensibles a les observacions anòmales. Per l'alt nombre d'observacions (respecte al nombre de paràmetres estimats) es pot despreciar la dependència entre residus (Peña 2000).

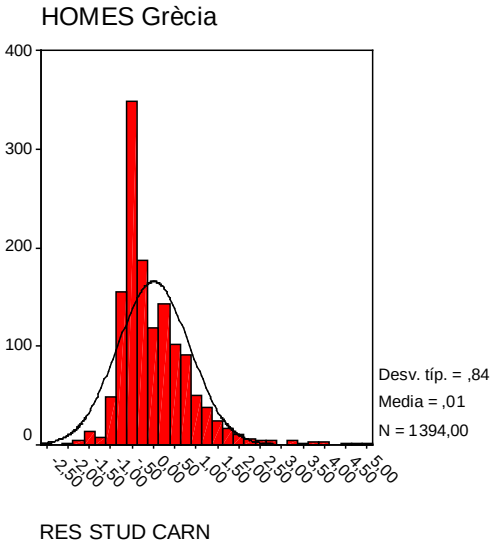
Per comprovar la hipòtesi de normalitat dibuixem un histograma dels residus estudentitzats i un dibuix de probabilitat normal. La majoria de centres tenen un comportament similar al dels homes d'Espanya (que es mostra), excepte Itàlia, el Regne Unit i Grècia (figura 11). En cada centre el comportament és similar entre homes i dones.

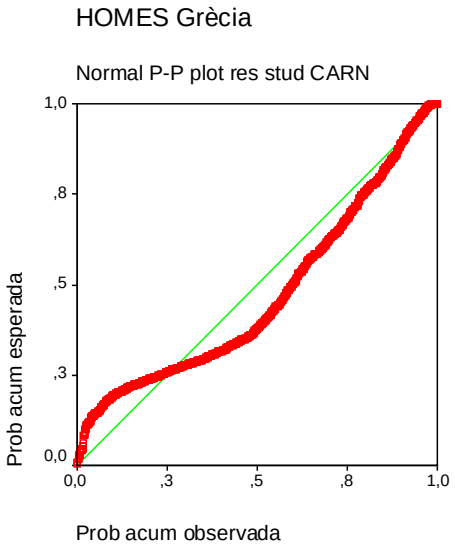
Figura 11. Histogrames i dibuixos de probabilitat normal pels residus estudentitzats del calibratge del consum de carn en homes d'Itàlia, Espanya, Regne Unit i Grècia.









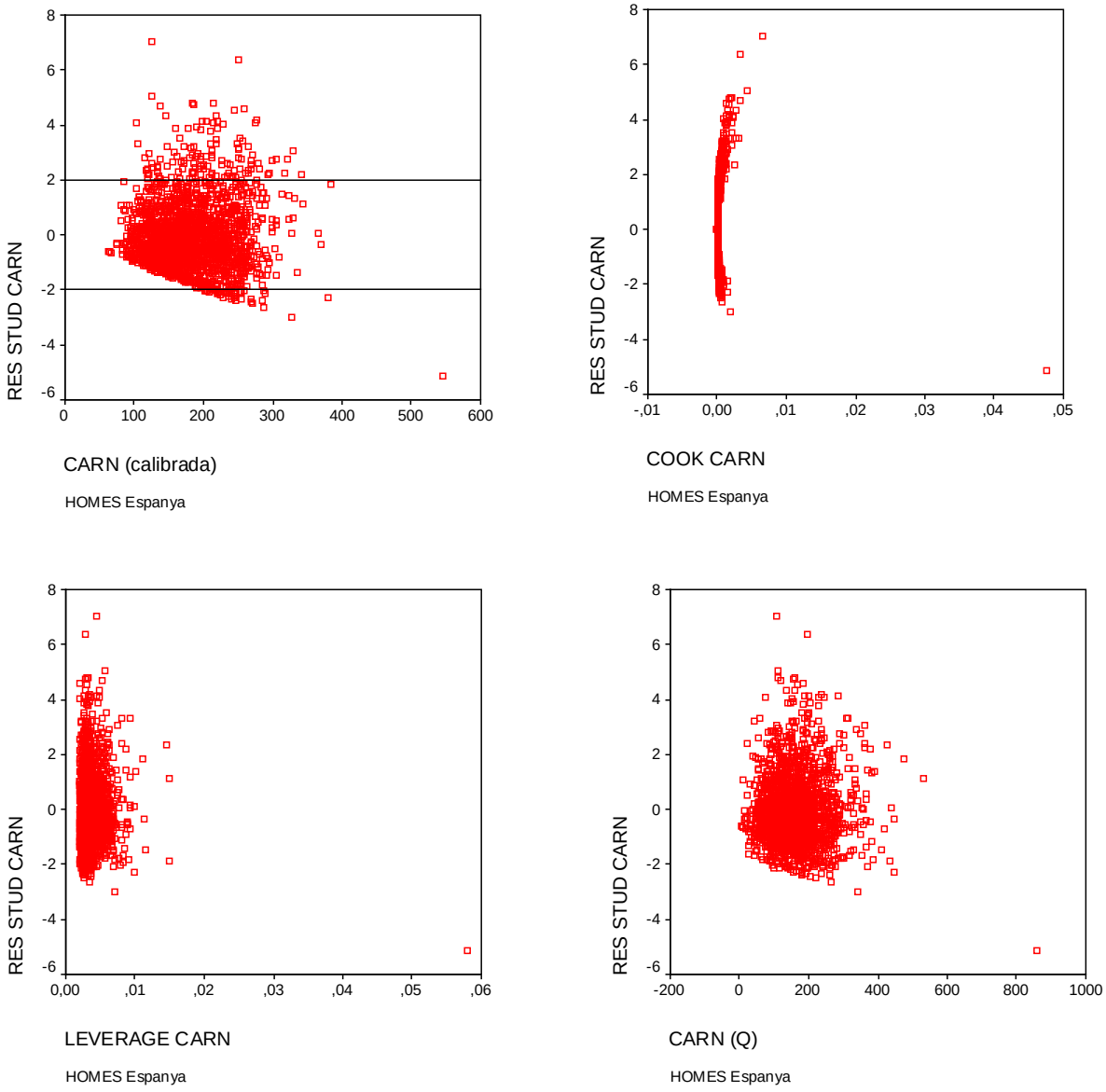


Vistos els gràfics anteriors és difícil d'assumir la normalitat dels residus, almenys en alguns centres com el Regne Unit o Grècia. L'efecte de la falta de normalitat és la pèrdua d'eficiència dels estimadors obtinguts. Per tant els tests sobre la significació del paràmetre poden ser no vàlids, però sí que podem limitar-nos a fer una estimació puntual del paràmetre (Peña 2000). Més endavant podem aproximar l'error estàndard del paràmetre mitjançant *bootstrap*.

El gràfic de dispersió dels residus estudentitzats envers la variable calibrada (o predita) es pot veure a continuació (els no consumidors en QFA/HD n'estan exclosos), per detectar la falta de linealitat, heterocedasticitat i valors atípics. També podem observar

els diagrames de dispersió dels residus estandarditzats respecte el coeficients de Cook i el *leverage* per mirar observacions influents a *posteriori* i a *priori*, respectivament. Per últim, també es creuen els residus respecte a la variable original de consum, per veure si aquesta pot explicar la possible falta de linealitat o heterocedasticitat dels residus. Es mostren els gràfics corresponents als homes d'Espanya (figura 12), que s'assemblen força a la resta de centres (en cada centre també s'assemblen els gràfics dels homes i els de les dones).

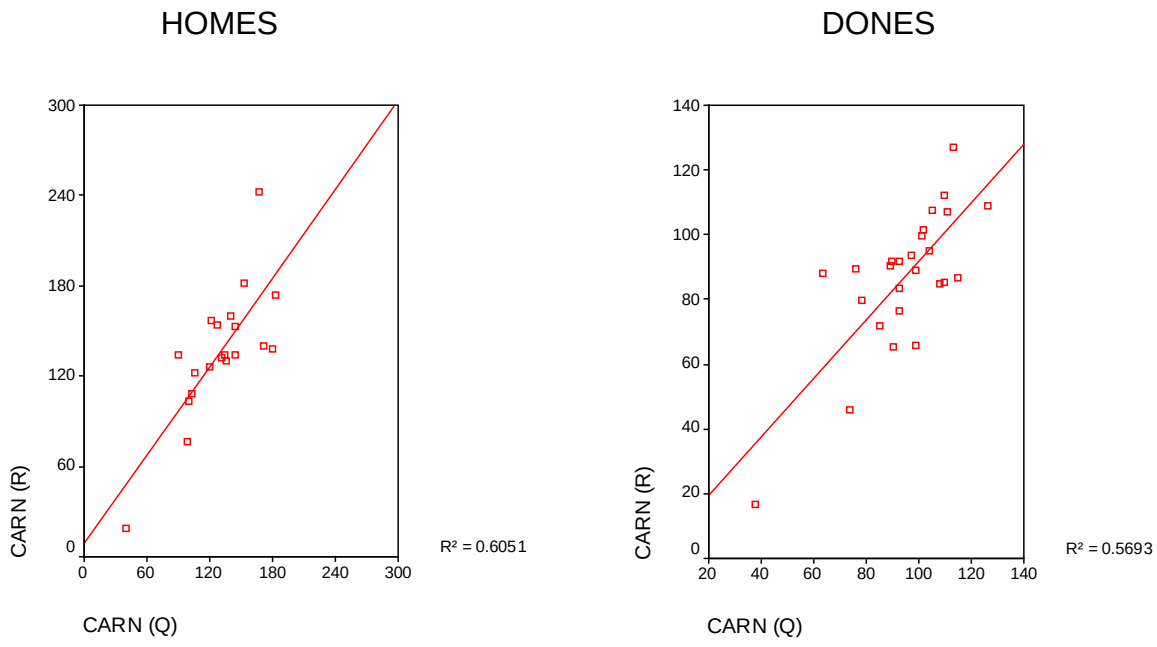
Figura 12. Diagrames de dispersió dels residus estudentitzats del calibratge de carn respecte la variable predita (calibrada), coeficient de Cook, *leverage* i consum de carn QFA/HD per als homes d'Espanya.



Es veuen uns residus sense forma aparent, però esbiaixats cap a valors positius (hi ha molts més residus per sobre de 2 que per sota de -2), degut probablement a l'asimetria dels valors extrems (poden haver-hi valors extrems només positius). S'observa una certa heterocedasticitat en alguns casos, però cal tenir en compte que la gran massa de punts queda centrada al voltant del 0. La recta descendent que es veu a la part baixa de la gràfica dels residus respecte la variable calibrada correspon als individus que no consumeixen carn en el R24H. Hi ha una sèrie reduïda de punts que podrien ser influents, com es desprèn de les gràfiques de Cook i *leverage*. Per últim, la relació de la variable original amb els residus és similar a l'observada amb la variable calibrada.

Com s'ha pogut observar, l'ajust dels models de calibratge no és gaire bo. Es podria mirar de millorar usant transformacions o eliminant certs punts influents. Recordem, però, que l'important es disposar d'una estimació puntual; no cal que sigui precisa, però sí que funcioni bé a nivell grupal. Observem, doncs, què passa si mirem la relació entre la variable original *Q* i la de referència *R* a nivell de centre per a homes i per a dones separadament (figura 13).

Figura 13. Diagrama de dispersió de les mitjanes de consum de carn per centre utilitzant el R24H (R) i el QFA/HD (Q).



Com es veu, l'ajust millora de forma espectacular, si bé la concordança entre ambdues mesures és heterogènia entre centres. L'ajust de formes quadràtiques o cúbiques només millora l'ajust lleument. Aquesta millora és deguda, en primer lloc, al fet que la mitjana dels R24H sí que és un bon mètode de referència i, en segon lloc, que la correlació entre mitjanes sempre serà major que la correlació individual degut a la disminució de l'error aleatori.

7.4 MODIFICACIONS AL MODEL DE CALIBRATGE ORIGINAL

A l'anex A2 es pot veure amb més detall què li passa al model de calibratge quan efectuem transformacions a la variable consum de carn, n'excloem els alts consumidors, fusionem els baixos consumidors amb els no consumidors o incloem l'energia com a variable d'ajust.

Si busquem la transformació de Box-Cox que ens normalitzi i faci més constant la variància de les dades obtenim valors de 0,65 i 0,64 per la carn del QFA/HD i 0,45 i 0,33 pel R24H per a homes i dones, respectivament. La transformació arrel quadrada millora lleument la normalitat dels residus, tot i que creuant la variable del QFA/HD amb la del R24H no s'aprecia cap millora important en els coeficients r^2. L'aplicació de logaritmes a les dades de consum tampoc suposa una millora en l'ajust del model de calibratge.

En una mostra gran, com la que tenim en l'estudi, és poc probable que uns pocs valors extrems modifiquin la relació entre variables. Repetint l'anàlisi excloent els qui consumeixen més del percentil 99 aproximadament, la variació en els coeficients de desatenuació respecte a l'ús de les variables sense exclusions és baixa. La distribució del consum de carn corresponent al QFA/HD és més simètrica i normal. Pels R24H el pes dels no consumidors continua essent molt important. Creuant les dues mesures excloent els grans consumidors no hi ha cap canvi evident en el núvol de punts ni en l'ajust de rectes de regressió. Els residus tenen un comportament gairebé idèntic al que tenien abans d'excloure els grans consumidors.

Poden aparèixer valors molt petits de consum provinents de receptes (habitualment receptes estàndard, que l'enquestat no pot modificar). Podem intentar reclassificar els consumidors de quantitats petites com a no consumidors. Hem triat consumidors habituals de menys de 3 grams al dia de carn com a "no consumidors". Els coeficients de desatenuació gairebé no varien al excloure els consumidors de menys de 3 g/dia i l'ajust del model de calibratge no varia gaire al excloure als baixos consumidors.

Alguns autors indiquen la necessitat d'ajustar pel consum total calòric (energia) a l'hora de calibrar. Afegint aquesta variable en el model de calibratge no es modifiquen pràcticament els resultats.

7.5 APLICACIÓ DE LES DADES CALIBRADES A UN MODEL DE COX PER CG

Els següents resultats es refereixen a la variable carn calibrada sense aplicar transformacions, exclusions o reclassificacions, si no s'esmenta el contrari. La variable és expressada de forma contínua, en consum de 100 grams/dia.

Inicialment es proven dos models separats per gènere per a la variable carn calibrada (per cada 100 grams de consum), estratificats per país i ajustats per centre. El model també s'ajusta per consum de tabac (mai fumador, ex-fumador i fumador actual), IMC (Kg/m^2), nivell educatiu (cap, primari, FP (tècnic), secundari, universitari i no especificat) i ingesta energètica (Kcal).

La variable energia no és significativa en cap dels dos models ($p>0,20$). El HR de patir CG varia d'1,80 a 1,50 en homes i de 3,24 a 2,96 en dones en deixar d'ajustar per energia. La correlació entre la variable calibrada i l'energia és de 0,40 en homes i 0,29 en dones, mentre que no trobem un efecte per l'energia ($p=0,47$ i $p=0,61$ respectivament per a homes i dones en un model univariat). Els models ajustats per energia simulen una situació en que l'energia es fixada però la composició de la dieta pot variar. Els requeriments energètics depenen de la mida corporal, activitat física i metabolisme dels individus, que poden confondre la relació entre la ingesta de l'aliment i la malaltia. L'ajust per energia pot reduir els efectes d'aquestes variables de confusió (Plummer 2003).

La variable nivell educatiu tampoc és gens significativa en dones, però en homes el p-valor global de la variable val 0,13 , observant-se una possible protecció en el cas d'homes sense estudis (HR universitaris respecte sense estudis=1,87 IC95%: 0,59-5,90), en desacord amb les evidències científiques publicades (van Loon 1998). El canvi al HR per consum de 100 grams de carn és d'1,80 a 1,81 i de 3,24 a 3,28 en excloure el nivell educatiu del model, en homes i dones respectivament. El consum (calibrat) de carn disminueix a mesura que augmenta el nivell educatiu (només en homes) (de 150 grams/dia per als que no tenen estudis fins a 125 grams/dia per als universitaris). La relació del nivell educatiu amb el CG és nul·la en dones ($p=1,00$) però el HR de la

majoria de categories respecte als homes sense estudis és al voltant de 2 (p=0,23). Per tant es podria tractar d'una variable de confusió, en homes, que cal incloure en el model.

La variable IMC cau a prop de la significació per als dos sexes (p=0,07 i 0,13 per a homes i dones respectivament). Cal dir, però, que el possible efecte protector observat desapareix en seleccionar els individus seguits més de dos anys (és probable que els individus perdin pes abans de ser diagnosticats). L'efecte del tabac ja és conegut (González 2003) i s'inclou com a variable d'ajust. Per tant, finalment no descartem cap de les 4 variables d'ajust inicialment proposades (tabac, escolaritat, IMC i energia).

Cal tenir també en compte l'efecte de la variable indicadora dels individus als quals se'ls ha assignat un valor calibrat de zero directament (els que tenien zero en la variable del qüestionari QFA/HD). La interpretació del coeficient per aquesta variable no és senzilla, ja que s'acompanya de la variable contínua calibrada. S'ha d'interpretar com el HR d'aquells que no consumeixen carn respecte als que tot i consumir-ne tenen un valor calibrat (predit) de zero. Ens permet tenir en compte en el model de malaltia que hem fet una assignació directa a una part de la mostra. Els valors dels HR per a aquesta variable indicadora són molt imprecisos, i en el cas de les dones és 0, ja que no hi ha cap malalta a qui s'hagi assignat zero directament. Els HR de carn varien d'1,80 a 1,64 en homes i de 3,24 a 3,32 en dones si excloem la variable indicadora.

La diferència dels HR de carn d'homes i dones és de més d'un punt (1,80 els homes i 3,24 les dones, només significativament diferent d'1 per a aquestes últimes) (taula 10). Provant un model amb la variable calibrada, les 4 variables d'ajust i la variable gènere, així com la interacció d'aquesta amb la variable calibrada i comparant-lo amb el mateix model sense el terme d'interacció, obtenim una diferència en la log-versemblança d'1,38 amb un grau de llibertat, cosa que indica que el model amb interacció no és significativament millor que el que no té aquest terme. Així doncs, el model que s'utilitzarà serà el conjunt per a homes i dones amb una variable indicadora que els diferenciï.

Els HR obtinguts amb el model conjunt (taula 11) se situen al mig dels obtinguts amb els models separats per gènere, però amb més significació, ja que es guanya potència en

Taula 10. *Hazard ratios* (HR) de patir CG, intervals de confiança al 95% (IC95%) i p-valors per models de Cox separats per gènere.

	Homes			Dones		
	HR	IC95%	p	HR	IC95%	p
Ex-fumador	1,21	0,79 1,85	0,388	1,61	1,03 2,51	0,037
Fumador actual	1,69	1,08 2,63	0,022	1,74	1,09 2,78	0,021
IMC (Kg/m^2)	0,95	0,90 1,00	0,069	0,97	0,92 1,01	0,132
Primària	2,37	0,81 7,00	0,117	1,09	0,41 2,90	0,860
FP	2,09	0,67 6,49	0,204	0,94	0,32 2,75	0,909
Secundària	1,29	0,38 4,32	0,685	1,02	0,35 2,93	0,979
Universitat	1,87	0,59 5,90	0,285	0,98	0,33 2,92	0,971
No especificat	0,38	0,04 3,88	0,411	1,07	0,22 5,15	0,930
Energia (Kcal)	1,00	1,00 1,00	0,214	1,00	1,00 1,00	0,652
CARN calibrada (x100 grams)	**1,80**	**0,98 3,31**	**0,060**	**3,24**	**1,30 8,07**	**0,012**
Zeros CARN	8,23	1,02 66,37	0,048	0,00	0,00 -	0,994

La referència per als ex-fumadors i fumadors actuals són els mai fumadors i pels nivells d'estudis primària, FP, secundària, universitat i no especificat els que no tenen estudis.

tenir més casos. El tabac continua essent un factor de risc (HR=1,34 i 1,72 per a ex-fumadors i fumadors actuals respectivament, p conjunta=0,003), mentre que no s'observa cap efecte global del nivell d'estudis (p conjunta=0,44). L'IMC apareix protector (HR=0,96, p=0,026) i el consum calòric no mostra cap efecte (HR=1,00, p=0,21). Les dones tenen menys risc de patir CG (HR=0,68, p=0,032). Per últim, la nostra variable d'interès, la carn calibrada, és un factor de risc de CG (HR=1,97, IC95%=1,21-3,22).

La millora aportada pel calibratge la podem estimar comparant els resultats obtinguts amb les variables calibrada i sense calibrar. Els resultats abans de calibrar no són gaire diferents (taula 12). Com hom podia esperar el valor del HR per a la variable sense calibrar és una mica menor (HR=1,43) i té un interval de confiança més estret (IC95%=1,13-1,81) (ja s'esperava que la variable calibrada donaria estimadors de HR menys precisos, ja que $Var(\hat{\beta}_i^*) = \frac{1}{\hat{\lambda}_i^2} Var(\hat{\beta}_i) + \frac{\hat{\beta}_i^{*2}}{\hat{\lambda}_i^4} Var(\hat{\lambda}_i)$ (i λ<1 i el segon terme s'acostuma a menysprear).

Veure que el paràmetre obtingut usant la variable corregida és com el que obtindríem dividint l'original per λ és molt difícil, tenint en compte que s'estratifica per país, es

pondera, etc. Tot i així, podem deduir una mena de λ global aïllant-la de $Var(\hat{\beta}_i^*) = \frac{1}{\hat{\lambda}_i^2} Var(\hat{\beta}_i)$, sabent que l'error estàndard del coeficient de HR per a la variable calibrada és 0,24965 i el de la variable original és 0,11918 , cosa que dóna una λ aproximada de 0,48. Si dividíssim el log(HR) obtingut amb el QFA/HD per 0,48 hauríem d'obtenir el log(HR) corregit en que es té en compte que el QFA/HD mesura el consum de carn amb error. En aquest cas, si fem β^*=log(1,43)/0,48 obtenim HR=exp(β^*)=2,11. Aquest valor hauria de coincidir amb l'1,97 que hem vist a la taula 11. En aquest cas no coincideix ja que la forma de calcular λ és aproximada.

Taula 11. *Hazard ratios* (HR) de patir CG, intervals de confiança al 95% (IC95%) i p-valors pel model de Cox conjunt per a homes i dones.

	HR	IC95%	p
Ex-fumador	1,34	0,99 1,82	0,056
Fumador actual	1,72	1,26 2,36	0,001
IMC (Kg/m^2)	0,96	0,93 1,00	0,026
Primària	1,51	0,77 2,95	0,230
Tècnica	1,33	0,64 2,75	0,441
Secundària	1,11	0,53 2,34	0,788
Universitat	1,25	0,60 2,61	0,548
No especificat	0,77	0,24 2,48	0,662
Energia (Kcal)	1,00	1,00 1,00	0,205
CARN calibrada (x100 grams)	**1,97**	**1,21 3,22**	**0,007**
Zeros CARN	2,83	0,38 21,13	0,311
Dones	0,68	0,48 0,97	0,032

La referència per als ex-fumadors i fumadors actuals són els mai fumadors i per als nivells d'estudis primària, FP, secundària, universitat i no especificat els que no tenen estudis.

Taula 12. *Hazard ratios* (HR) de patir CG, intervals de confiança al 95% (IC95%) i p-valors pel model de Cox conjunt per a homes i dones usant la variable sense calibrar.

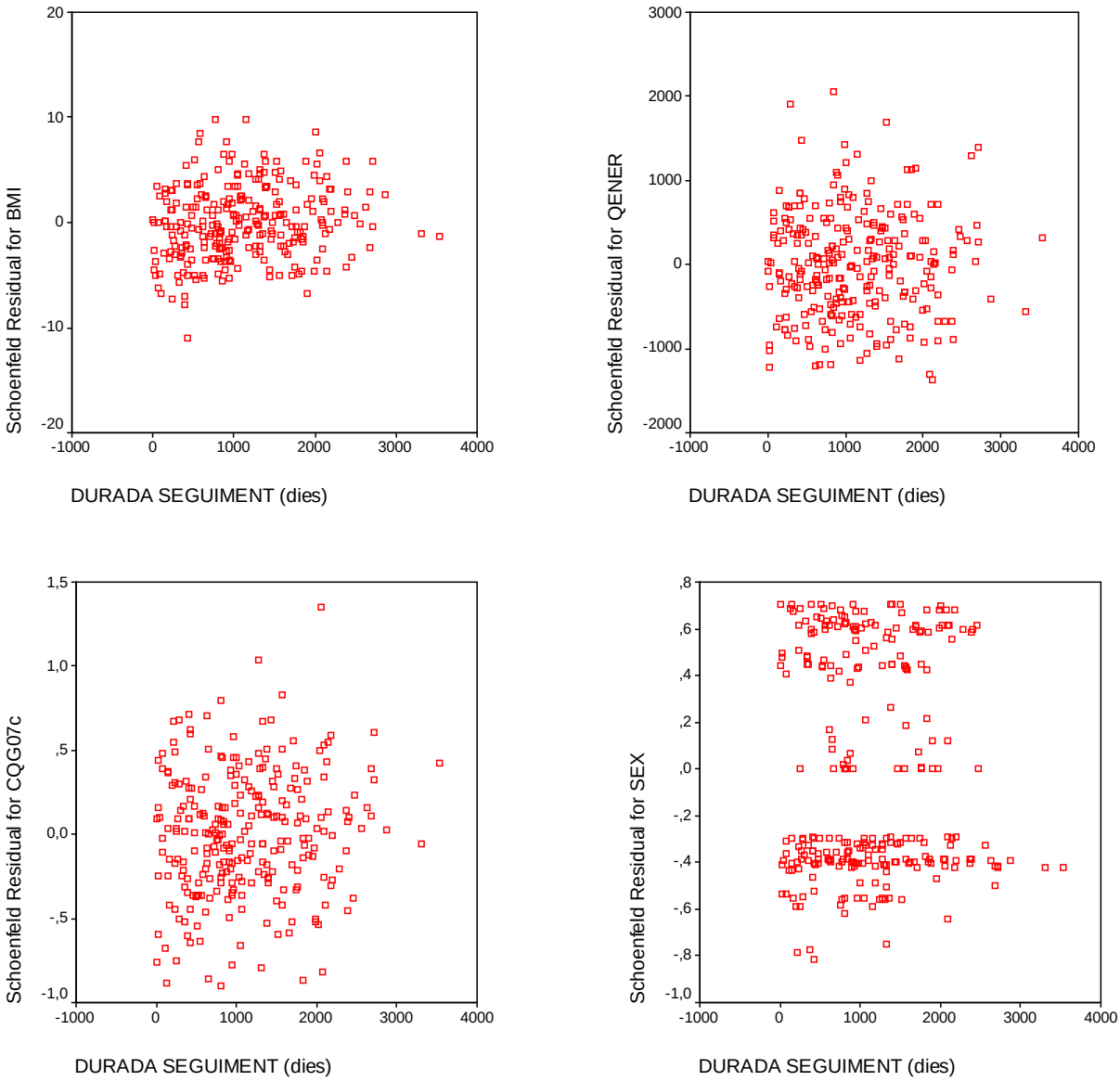
	HR	IC95%		p
Ex-fumador	1,35	1,00	1,83	0,054
Fumador actual	1,72	1,26	2,36	0,001
IMC (Kg/m^2)	0,97	0,94	1,00	0,048
Primària	1,51	0,77	2,95	0,231
Tècnica	1,34	0,65	2,77	0,431
Secundària	1,11	0,53	2,35	0,778
Universitat	1,27	0,61	2,64	0,526
No especificat	0,78	0,24	2,50	0,673
Energia (Kcal)	1,00	1,00	1,00	0,147
CARN original (x100 grams)	1,43	1,13	1,81	0,003
Dones	0,55	0,40	0,74	<,0001

La referència per als ex-fumadors i fumadors actuals són els mai fumadors i per als nivells d'estudis primària, FP, secundària, universitat i no especificat els que no tenen estudis.

7.6 AJUST DEL MODEL DE COX

Un cop hem triat el model de Cox, cal comprovar-ne la bondat de l'ajust. Els residus de Schoenfeld, específics per a cada variable predictora, poden ajudar a detectar variables que no compleixin el supòsit de proporcionalitat. A continuació podem veure els diagrames de dispersió dels residus de Schoenfeld respecte el temps de seguiment per a algunes variables predictores (figura 14).

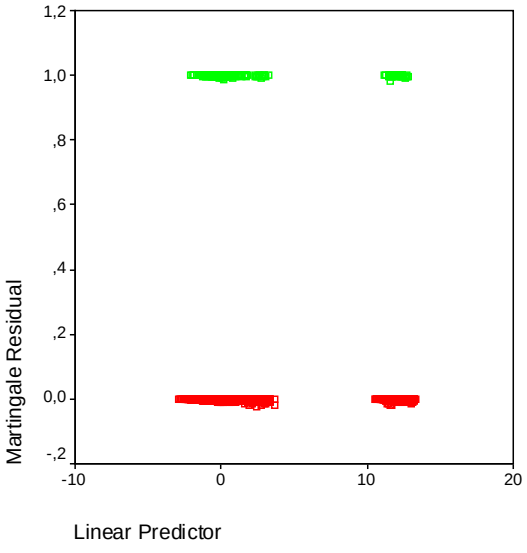
Figura 14. Diagrames de dispersió dels residus de Schoenfeld dels casos per IMC, energia, consum de carn calibrada (x 100 grams) i gènere respecte el temps de seguiment.



Com s'ha vist ens els gràfics anteriors, no es detecta cap tendència identificable (en la resta de variables no dibuixades tampoc es mostra cap tendència). Les variables dicotòmiques separen els residus en dos grups, però no s'observa tendència a cap dels grups d'aquestes variables. A més, fent el test $\beta_j(t) = \beta \quad \forall t$ (on t és el temps de seguiment) mitjançant una khi-quadrat basat en els residus de Schoenfeld (Stata 2003) obtenim p=0,91. Per tant, res s'oposa a assumir proporcionalitat en els riscos.

A la figura 15, en què es creua el predictor lineal (resultant d'emprar totes les covariables del model) amb els residus de martingala, veiem que el model no classifica bé els casos, ja que s'espera que els residus de martingala estiguin al voltant de zero. Això és degut al fet què el nombre de casos (color verd) és ínfim comparat al d'individus a risc (vermell) i per tant l'error global de classificació és molt baix tot i classificar malament a tots els casos. El salt que s'observa per al predictor lineal és degut al fet que el model ajusta els paràmetres dels marcadors dels centres italians respecte a Nàpols, que no té cap cas, cosa que dóna estimacions molt altes, però això no ha de representar cap problema en l'estimació dels coeficients en que estem interessats (bàsicament el de la carn).

Figura 15. Diagrama de dispersió dels residus de martingala respecte el predictor lineal. Casos en color verd i individus a risc en color vermell.



Repetint el mateix gràfic per a cadascuna de les variables predictores individualment el resultat és anàleg.

Recordem, però, que estem més interessats en buscar una associació entre CG i carn que en fer prediccions. Per tant, una hipòtesi a comprovar és la linealitat de l'efecte. Repetint el mateix gràfic anterior dels residus de martingala, però ara respecte el consum de carn, si fem un suavitzat del diagrama de dispersió esperarem trobar una corba aproximadament plana respecte al consum de carn. Si no trobem aquesta corba voldrà dir que la forma funcional de la variable independent (carn) no és acceptable. Com es veu a la figura 16, obtenim una línia gairebé recta a l'alçada dels individus a risc, que ens permet acceptar l'assumpció de linealitat de l'efecte de la carn calibrada.

Una anàlisi dels punts influents, mitjançant l'estadístic LD (desplaçament de la versemblança en eliminar un punt) (SAS 2001) mostra cinc casos com a possibles punts influents (figura 17).

Figura 16. Diagrama de dispersió dels residus de martingala respecte el consum de carn calibrada (x 100 grams) i suavitzat (línia negra) del diagrama.

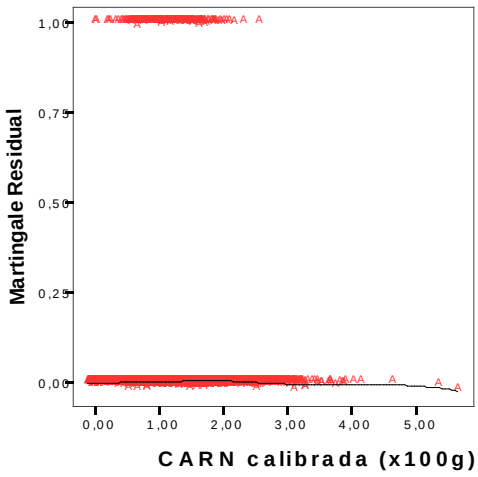
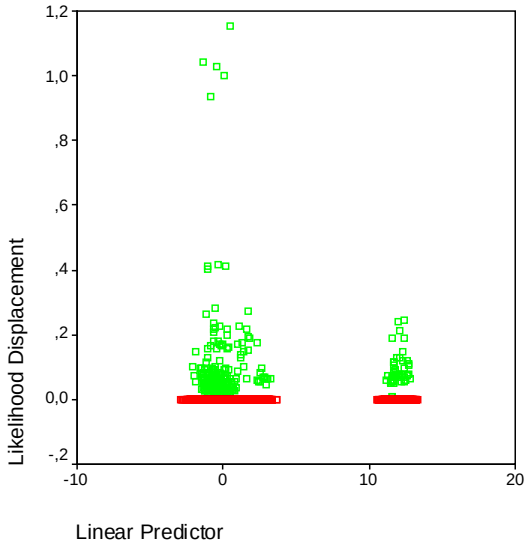


Figura 17. Diagrama de dispersió de l'estadístic LD respecte al predictor lineal. Casos en color verd i individus a risc en color vermell.



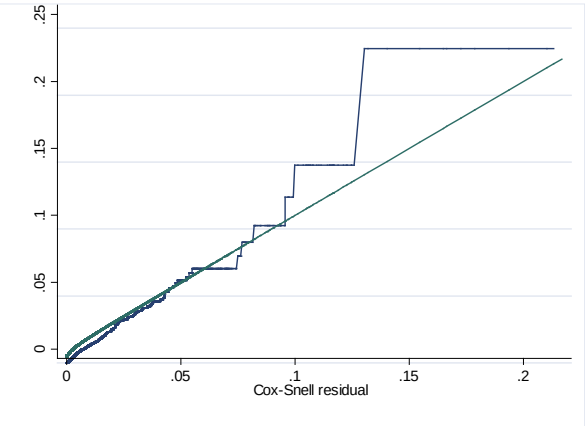
Eliminant aquestes cinc observacions es modifiquen lleugerament els coeficients obtinguts pel model de Cox (el HR per 100 grams de carn passa d'1,97 a 1,93 , en tots dos casos significativament diferent d'1 (p=0,007 i p=0,010 pel model amb tots els casos i amb l'exclusió dels presumptes punts influents respectivament).

L'estratificació per edat, a més de per país, no aporta grans canvis, probablement perquè l'estratificació per país ja té en compte les diferències en l'estructura d'edat. Concretament el HR per carn passa d'1,97 (p=0,007) a 2,01 (p=0,052) en estratificar per aquesta variable (categoritzada com <45, 45-55, 55-65 i >65 anys). Cal dir que en estratificar per país i edat alguns grups deixen de tenir casos, deixant de figurar al denominador tot l'estrat. Per tant, és preferible no estratificar per edat per aquest fet i per la similitud dels resultats.

Una forma alternativa de veure l'ajust global del model és mitjançant els residus de Cox-Snell. Aquests residus han de tenir una distribució exponencial si el model de Cox ajusta bé les dades. S'espera que en una gràfica en què es creuin els residus de Cox-Snell amb el risc acumulat calculat usant els residus de Cox-Snell com a eix de temps i l'event original (CG) com a variable de censura, s'obtingui una recta de pendent 1 (Stata 2003). Això és el que obtenim amb les nostres dades, excepte a la part dreta de la

gràfica, en què hi ha molts pocs casos i la funció de risc basal es fa molt inestable (figura 18).

Figura 18. Residus de Cox-Snell.



Per tant, podem concloure que l'ajust del model utilitzat és raonablement acceptable.

7.6.1 HOMOGENEÏTAT

En estratificar per país assumim que entre aquests hi ha un cert efecte homogeni del consum de carn sobre el CG. Si bé observem a la taula 13 diferències entre països (el HR per 100 grams de carn varia de 0,83 a Espanya a 2,17 a França si usem la variable original i de 0,42 a Grècia a 13,60 a França si usem la variable calibrada), els estimadors són molt imprecisos, cosa que fa que en fer un test d'heterogeneïtat no obtinguem resultats significatius (obtenim una khi-quadrat amb 8 graus de llibertat de 7,25 per a les dades originals i de 6,22 per a les dades calibrades, en comparar la versemblança entre models amb termes d'interacció país-consum de carn i sense).

Taula 13. *Hazard ratios* (HR) de patir CG, intervals de confiança al 95% (IC95%) i p-valors per models de Cox per a la variable carn (x100 grams) calibrada i sense calibrar per país. Models amb homes i dones junts.

	Calibrat			Original		
	HR	IC95%	p	HR	IC95%	p
França	13,60	0,32 582,30	0,173	2,17	0,69 6,78	0,183
Itàlia	4,98	1,27 19,50	0,021	2,11	1,18 3,77	0,012
Espanya	0,71	0,11 4,47	0,715	0,83	0,37 1,84	0,640
Regne Unit	2,95	1,46 5,95	0,003	1,72	1,19 2,48	0,004
Holanda	1,41	0,27 7,29	0,684	1,27	0,46 3,56	0,644
Grècia	0,42	0,01 13,33	0,624	0,93	0,17 4,93	0,927
Alemanya	0,87	0,32 2,41	0,795	0,96	0,55 1,69	0,893
Suècia	1,50	0,30 7,49	0,621	1,31	0,71 2,45	0,390
Dinamarca	5,42	0,20 149,49	0,318	1,78	0,70 4,51	0,225

7.6.2 SEGUIMENT DE MÉS DE 2 ANYS

Un dels possibles problemes quan s'avalua la relació entre el consum de carn i el CG és que els individus als quals se'ls ha diagnosticat un càncer en els primers mesos de seguiment podrien haver canviat els seus hàbits dietètics (i altres, com el consum de tabac) precisament perquè ja tenien la malaltia (o una precursora d'aquesta) però encara no havien estat diagnosticats.

El consum diari de carn en els diagnosticats de CG en els primers dos anys és de 110 grams i en els diagnosticats després de 115 grams. Repetint l'anàlisi, excloent-ne els casos i individus a risc que tenen un seguiment menor de 2 anys (resten 180 casos i 440.121 individus a risc disponibles per l'anàlisi) el HR per carn (x100 g.) creix fins 2,30 (IC95%: 1,29-4,11), desapareix l'efecte de l'IMC (p=0,52) i augmenta el del tabac (HR per a fumadors actuals passa de 1,72 a 2,04) (taula 14).

Taula 14. *Hazard ratios* (HR) de patir CG, intervals de confiança al 95% (IC95%) i p-valors per models de Cox per a la variable carn (x100 grams) excloent els seguits durant menys de 2 anys.

	HR	IC95%	p
Ex-fumador	1,45	1,00 2,11	0,053
Fumador actual	2,04	1,39 3,00	0,000
IMC (Kg/m^2)	0,99	0,95 1,03	0,522
Primària	1,21	0,56 2,59	0,632
Tècnica	0,99	0,43 2,28	0,973
Secundària	0,95	0,40 2,23	0,902
Universitat	1,04	0,45 2,41	0,931
No especificat	0,94	0,25 3,52	0,923
Energia (Kcal)	1,00	1,00 1,00	0,245
CARN calibrada (x100 grams)	**2,30**	**1,29 4,11**	**0,005**
Zeros CARN	0,00	0,00 -	0,989
Dones	0,73	0,48 1,13	0,158

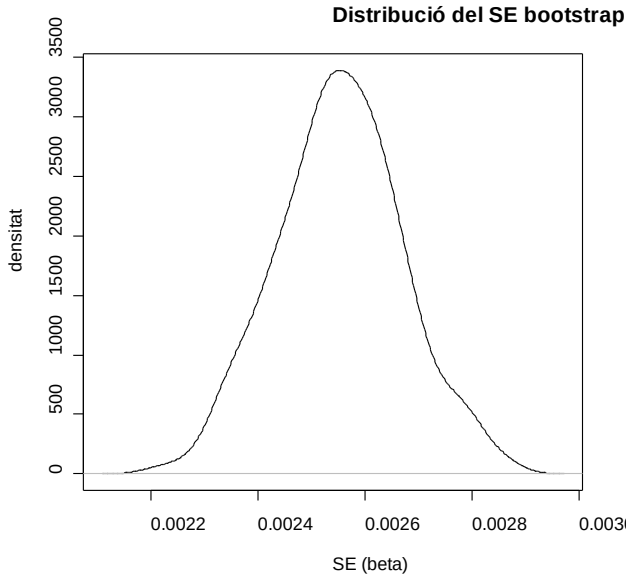
La referència per als ex-fumadors i fumadors actuals són els mai fumadors i per als nivells d'estudis primària, FP, secundària, universitat i no especificat els que no tenen estudis.

7.6.3 CORRECCIÓ DE LA VARIÀNCIA

Cal recordar que fins ara no s'ha tingut en compte que la variable que s'usa en el model de Cox és una predicció d'un model de regressió lineal. Per tant, la variància estimada està infraestimada, ja que no té en compte la variabilitat que el model de calibratge aporta. Per resoldre aquesta situació es fa un procediment *bootstrap*, que calcula 300 vegades el HR del model de Cox, a partir de 300 estimacions de la variable carn calibrada. S'usa el nombre total d'observacions amb qüestionari basal i de referència, però els individus s'escullen de forma aleatòria amb repetició, cosa que provoca que algun individu pugui estar repetit o algun altre exclòs en el model de calibratge (que com s'ha dit s'executa 300 cops). Un cop disposem de les 300 estimacions del paràmetre central (log(HR)) i 300 estimacions de la seva variància es pot calcular la variància corregida a partir de la mitjana de les 300 variàncies estimades i de la variància de les 300 estimacions del log(HR). L'error estàndard sense fer *bootstrap* era de 0,0025 grams/dia. L'error estàndard corregit val 0,002621 grams/dia, o sigui un 4,84% més. D'aquesta forma els intervals de confiança originals del HR (1,21-3,22) es corregirien a (1,18-3,30). En la figura 19 podem veure la distribució dels 300 errors estàndard de beta [=log(HR)] calculats. La línia verda indica l'error estàndard del

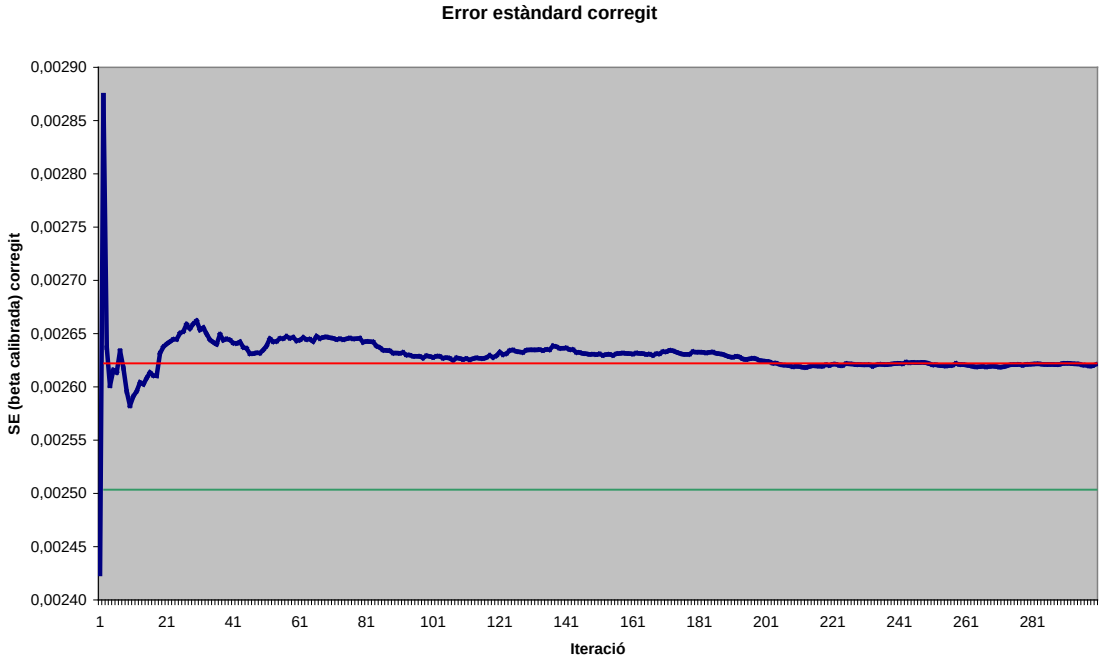
log(HR) amb el mètode no *bootstrap* i la vermella la mitjana dels 300 errors estàndard estimats.

Figura 19. Distribució dels errors estàndard (SE) del log(HR) de carn calibrada (x 1 g.) calculats en cada iteració del *bootstrap*. Línia verda: SE original sense usar *bootstrap.* Línia vermella: mitjana dels 300 SE obtinguts en cada iteració del *bootstrap.*



També podem observar com evoluciona l'error estàndard corregit amb cada iteració. Amb 200 iteracions ja n'hi hauria hagut prou per tenir una bona estimació. La línia vermella indica l'error estàndard corregit estimat per *bootstrap* i la verda l'error estàndard que havíem obtingut sense fer *bootstrap* (figura 20).

Figura 20. Evolució de l'error estàndard (SE) corregit del log(HR) de carn calibrada (x 1 g.) en cada iteració del *bootstrap* (línia blava). Línia vermella: SE corregit obtingut després de 300 iteracions *bootstrap*. Línia verda: SE no corregit del model original.



7.6.4 MODELS AMB VARIABLES TRANSFORMADES, EXCLUSIONS O RECLASSIFICACIONS

Anteriorment hem justificat perquè no calia modificar la variable carn usada per calibrar. Malgrat això, podem veure a la taula 15 els HR obtinguts, els seus intervals de confiança i el p-valor usant el model escollit per a l'anàlisi i altres models en què s'usa la variable transformada o amb restriccions, tant per calibrar com pel model de risc.

En qualsevol dels casos sempre s'observa un efecte de risc per la carn. Els resultats del model original no varien al ajustar per energia en el calibratge ni agrupant els consumidors de menys de 3 grams de carn al dia amb els no consumidors. La

Taula 15. *Hazard ratios* (HR) de patir CG per consum de carn, intervals de confiança al 95% (IC95%) i p-valors per a diferents models de Cox: usant la variable carn calibrada sense transformar, la transformació arrel quadrada i logaritme en base 2, excloent-ne els grans consumidors de QFA/HD i R24H, reclassificant els consumidors de menys de 3 grams/dia en QFA/HD com a no consumidors i ajustant per energia en el model de calibratge.

MODEL	HR	IC95%	p-valor
Contínua (x 100g)	1,97	1,21 - 3,22	0,007
Arrel (x 1g)	1,17	1,07 - 1,28	0,001
Log-2 (x 1g)	1,27	1,09 - 1,48	0,002
Excloent P99 (x 100g)	2,47	1,40 - 4,35	0,002
Considerant <3 g/dia no consumidor (x100g)	1,97	1,20 - 3,25	0,008
Ajustant per energia al calibrar (x100g)	1,97	1,21 - 3,20	0,006

comparació amb els models en què s'usa la transformació per arrel quadrada o per logaritme en base dos és més difícil per venir expressats en escales diferents. De qualsevol forma la significació és similar. En el cas del logaritme ja havíem vist que el comportament dels seus residus era el més anòmal. Per últim, si excloem els consumidors per sobre del percentil 99 (P99) el canvi és considerable. Hem vist que el model escollit ajustava prou bé i que l'exclusió de punts influents i veritables *outliers* no modificava gaire els resultats del calibratge. L'exclusió d'un u percent de la mostra suposa excloure més de 4.600 individus, entre ells només un cas. En realitat a aquests 4.600 individus no se'ls pot definir com a *outliers* sinó com a grans consumidors. Per tant, la seva exclusió hauria de venir suportada per motius nutricionals o biològics a part d'estadístics. Tot i així, pot indicar una certa infravaloració de l'efecte detectat en el model seleccionat.

8. DISCUSSIÓ

Si comparem els nostres resultats amb els d'altres estudis, els valors d'r^2 del model de calibratge s'assemblen als obtinguts en uns altres estudis usant el R24H com a eina de referència. Per exemple, Rosner (2001) obtingué un r^2 per carn de 0,06 mentre que aquí observem valors entre 0,02 i 0,18 (segons país). No és clar que la carn en general sigui un factor de risc per CG si ens atenem a la literatura (WCRF 1997). En qualsevol cas les mateixes fonts sí que citen alguns tipus de coccions i preparacions (embotits, carn fumada o curada) de la carn com a factors de risc de CG. Ward (1997) troba un OR de 2,4 per als qui consumien més de 19 cops a la setmana carn vermella envers els que en consumien menys de 8 cops. Així, és probable que el HR de 1,97 que hem trobat pugui ser més gran si ens centrem en alguns tipus concrets de carn o en com es prepara, conserva o cuina.

8.1 ASSUMPCIONS DEL MODEL

A continuació es discuteixen cadascuna de les assumpcions en què es basa la teoria del calibratge i l'aplicació de la variable calibrada obtinguda en un model de Cox.

no esbiaixada: si la mesura de la dieta *R* en el subestudi de calibratge és no esbiaixada (respecte a la realitat *T*) es poden obtenir estimadors no esbiaixats de λ usant una mesura única de la dieta (com el R24H) (Rosner 1990). Aquesta reducció del nombre de mesures, però, comportarà un increment dels errors estàndard de λ i per tant uns intervals de confiança més amples per β*. Una solució seria, aleshores, incrementar el nombre de subjectes participants a l'estudi de calibratge (Rosner 1988). La participació en l'estudi de calibratge ha de ser alta per evitar biaixos de selecció indesitjables. En l'estudi EPIC, com hem vist, la participació va ser prou alta (superior al 75% en 7 països). En aquest projecte no és pot demostrar si la mesura de referència és o no esbiaixada en no disposar d'una tercera mesura de referència no correlacionada amb els errors de *R* i *Q* i no esbiaixada (marcador bioquímic). Però, fins i tot si *R* fos esbiaixada però la direcció i magnitud de l'error fos aproximadament igual per les

diferents subcohorts, la mesura de referència encara es podria usar per fer un calibratge a nivell ecològic (calibratge entre cohorts o estandarditzar les mesures per obtenir un estimador comú). Això vol dir que les mesures calibrades no tindrien validesa absoluta, però si relativa. Fins i tot seria vàlida per corregir a nivell individual si l'error fos homogeni dins de cada grup (Kaaks 1997, Riboli 2000).

Slimani (2003) recorda que en absència d'un mètode de referència totalment lliure d'error, el calibratge dels *Q* encara permet la millora de la comparabilitat de les mesures entre diverses cohorts si l'error sistemàtic de l'instrument de referència és relativament modest i constant entre les diverses cohorts i el mètode de referència permet caracteritzar acuradament el consum mig en cada centre. Dins de l'EPIC comparen els resultats de proteïna i energia a partir de la concentració de nitrogen en orina, el R24H i el QFA/HD. La correlació ponderada per la mida de cada centre entre el biomarcador i el R24H és 0,83 i 0,95 i amb el QFA/HD 0,53 i 0,86 pel nitrogen i l'energia respectivament, essent més altes en homes que en dones, tant pel R24H com pel QFA/HD i similar entre països pel R24H. Aquests resultats indicarien, almenys pel nitrogen, que si bé la ingesta dels R24H o QFA/HD no està exempta d'error en termes absoluts, l'ordre i magnitud del biaix sistemàtic (habitualment infraestimació) del mètode de referència és, en general, comparable entre centres, i que en absència d'un mètode de referència completament lliure de biaix, una infra o supraestimació pot permetre encara reescalar les mitjanes de cada grup obtingudes amb els QFA/HD (Riboli 2002). Els errors del QFA/HD, al contrari que pel R24H, eren de diferent magnitud entre els centres i de diferent direcció, sobretot en dones (el que fa encara més necessari el calibratge (separat per sexes)). Altres autors han demostrat com les dones tendeixen a reportar menys energia que els homes (controlant per pes, altura i edat) (Black 1996) i a reportar pitjor la dieta (Ferrari 2002). En el nostre estudi s'observa com els homes tenen coeficients de calibratge superiors a les dones, el que podria reafirmar la hipòtesi de que les dones reporten pitjor el seu consum.

Per últim, recordar que si *R* és no esbiaixada, automàticament es compleixen dues assumpcions més: que ε_R té mitjana zero i que és independent de *T*.

Independència entre els errors de i (donat *T*): és difícil d'assumir en general (Kaaks 1995a) i impossible de provar en aquest projecte. Qüestionaris en què s'usa la

memòria de l'individu per contestar poden tenir errors correlacionats, sobretot si s'administren en períodes de temps molt pròxims (Freedman 1991). Si això passa, la variància del consum estimat *X* pot estar sobreestimada. En l'EPIC els dos qüestionaris utilitzats es basen en respostes a preguntes, per tant s'utilitza la memòria, i existeix heterogeneïtat quant al temps transcorregut entre la mesura del R24H i el QFA/HD en la mostra de calibratge. Una possible solució seria l'ús de marcadors bioquímics com a mètode de referència, però lamentablement no es disposa d'un marcador per a cada variable nutricional (Hunter 1990). Uns altres estudis (Cameron 1988) diuen que els mecanismes memorístics per recordar el consum del darrer dia (com el R24H) o a llarg termini (com QFA/HD) són diferents, fet que implicaria una menor correlació entre els errors. Si la correlació entre els errors és negativa l'estimador de RR (o HR) pot estar sobrecorregit (per sobre del valor real) (Wacholder 1993). Tot i així sembla força improbable que els errors de dos mètodes de mesura basats en qüestionaris tinguin correlació negativa. Spiegelman (1997a) demostra com la no correlació entre els errors d'*R* i *Q* dóna lloc a estimadors no esbiaixats de RR (suposant que e_R té mitjana 0 i cov(e_R,T)=0). També mostra com el biaix relatiu de l'estimador corregit del RR varia amb la correlació entre els errors d'*R* i *Q* , amb la fiabilitat d'*R* (defineix fiabilitat=var(R)/var (T)) i amb la correlació entre *R* i *Q*. Concretament l'error d'atenuació del RR (tendència cap a RR=1) s'incrementa a mesura que incrementem la corr(e_R,e_Q), o quan disminueix la fiabilitat d'*R* o quan disminueix la qualitat de *Q* (disminueix corr (Q, T)). Quan la fiabilitat d'*R* és 100% o la corr(e_R,e_Q)=0 o corr(Q,T)=1 no hi ha biaix en l'estimador de RR. Es veu que sempre que corr(e_R,e_Q) no sigui negativa és millor calibrar. Spiegelman (1997a) continua dient que amb una tercera mesura en el subestudi de calibratge (biomarcador *L*) amb errors no correlacionats amb e_R ni e_Q ni amb *T*, i amb rèpliques en *R*, podríem estimar corr(e_R,e_Q). Concretament, el nou factor de calibratge seria

$$\hat{\lambda}_{LRQ} = \frac{\hat{\lambda}_{L|Q}}{\hat{\lambda}_{L|R}} \frac{\hat{Var}(T)}{Var(R)} \qquad (36)$$

amb $\hat{\lambda}_{L|Q}$ i $\hat{\lambda}_{L|R}$ pendents de la regressió de *L* amb *Q* i *R* respectivament, i

$$\hat{Var}(T) = Var(R) - \hat{Var}(e_R) \qquad (37)$$

amb

$$\hat{Var}(e_R) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N_i} (R_{ij} - \bar{R}_i)^2}{(\sum_{i=1}^{N} N_i) - N} \quad (38)$$

on N és el nombre total d'individus a l'estudi de calibratge, i N_i el nombre de rèpliques d'R de l'individu i. Com sempre, l'estimador de risc relatiu corregit serà

$$\hat{\beta}^* = \hat{\beta} / \hat{\lambda}_{LRQ} \quad (39).$$

Spiegelman (1997a) dóna estimadors de la variància (que serà més gran que quan usàvem només R i Q per calibrar). Això indica que si poguéssim fer servir un tercer mètode de mesura apropiat, només hauríem d'emprar aquest coeficient de calibratge si sabem que hi ha correlació entre els errors de Q i R, perquè paguem el preu d'incrementar la variància.

Kipnis (2001) divideix el biaix en dues parts: el que depèn de la ingesta real (biaix de grup, que no existirà si es compleix l'assumpció que els errors d'R són no correlacionats amb T) i el que depèn de l'individu (no existirà si es compleix que els errors d'R i Q són no correlacionats). En el mateix article es diu que cal ajustar per estació (Landin 1995), tal i com nosaltres fem. Una forma fàcil de definir la correlació entre els errors (Kipnis 2001) és $\rho(\varepsilon_Q,\varepsilon_R)$ amb

$$Q=\phi_Q+\delta_Q T+\varepsilon_Q \quad (40)$$

$$R=\phi_R+\delta_R T+\varepsilon_R \quad (41)$$

Kipnis (2001) critica el model de Rosner (per tenir massa assumpcions, la vulneració de les quals rebaixaria la potència) però en la proposta que fa calen mesures repetides. Kipnis (1999) divideix l'error e_Q en sistemàtic (que depèn de T, i per tant seria el mateix per a tots els individus que consumeixen la mateixa quantitat real) i biaix individual (degut a característiques individuals). Els seus models no assumeixen la independència

entre e_Q i e_R (ni en la part sistemàtica ni en la individual) si s'administren al mateix temps, tot i que segueix assumint que *R* no té biaix. El problema és que quan arribem a aquests nivells de detall i complexitat el model no és identificable (més paràmetres que equacions). Per tant tot es basa en anàlisis de sensibilitat, assumint diversos valors per a $\rho(\varepsilon_Q,\varepsilon_R)$ i/o en la incorporació de noves mesures (biomarcadors). Conclou que si s'ignora el biaix individual en l'instrument de mesura, el coeficient de desatenuació estarà més a prop d'1 que allò que realment val (o sigui, no es té en compte tota l'atenuació). Per tant, en qualsevol cas l'estimador obtingut en calibrar en aquest projecte serà més proper al real que si no calibrem, si bé la correcció efectuada pot ser insuficient per arribar al coeficient real.

Variable contínua: la variable considerada (consum de carn) es pot considerar contínua, ja que no s'aprecien agrupacions de valors que la facin considerar discreta.

Relació lineal: no s'observa una relació lineal entre *Q* i *R* (s'usa *R* en no disposar de *T*): els r^2 són sempre menors que 0,20. Però ja sabem que *R* no és un bon estimador de *T* a nivell individual sinó a nivell grupal (país). En aquest cas r^2 puja a 0,61 en homes i 0,57 en dones. Regressions quadràtiques i cúbiques aporten millores poc importants (per exemple, r^2 de 0,65 i 0,66 respectivament per regressió quadràtica i cúbica en homes).

Normalitat: la normalitat conjunta del consum real i del consum mesurat no sembla ser tan assumible, ja que moltes mesures dietètiques tenen cues que s'aproximen a una distribució log-normal, cosa que provocaria un major error aleatori per consums alts. L'ús de models no paramètrics (o semiparamètrics) pot resoldre el problema (Carroll 1991). Aquest mateix autor arriba a dir que l'assumpció de normalitat no és necessària (Carroll 1990). Depèn, però, de la variable considerada. Per exemple, el percentatge energètic provinent del greix o transformacions logarítmiques o amb arrel quadrada del consum de nutrients poden ser normals (Carroll 1996). Almenys caldria que la distribució fos simètrica. Kuha (1994) diu que el requeriment principal perquè $\hat{\beta}^*$ sigui no esbiaixat és que p(D|T) sigui baix i f(T|X) normal. En aquest projecte la normalitat dels residus observats és discutible, sobretot en alguns centres com el Regne Unit o Grècia. Com veurem més endavant, l'ús de transformacions per millorar la normalitat

suposa incomplir una altra assumpció: E[R]=T (Boshuizen 2004). L'efecte de la falta de normalitat és la pèrdua d'eficiència dels estimadors obtinguts. Per tant els tests sobre la significació del paràmetre poden ser no vàlids, però sí que podem limitar-nos a fer una estimació puntual del paràmetre (Peña 2000). Més endavant podem aproximar l'error estàndard del paràmetre mitjançant *bootstrap*. La normalitat de la ingesta real o mesurada no sembla tan important, com diu Carroll (1990). En qualsevol cas, les mesures observades en aquest qüestionari no són normals, ja que tenen una important cua per la dreta (Q) i un elevat percentatge de no consumidors (R).

Independència dels errors de mesura d'altres característiques dels individus: els errors de mesura existents han de ser independents entre ells, de la ingesta real i d'altres característiques dels individus (Prentice 1996). Això no és gaire clar per variables com l'IMC. Usant mètodes de laboratori com a referència, s'ha trobat que les persones obeses (quart quartil d'IMC) infrareporten fins un 30-40% de l'energia consumida respecte a les del primer quartil d'IMC (Heittman 1995). En canvi, els mateixos autors troben menys diferències entre els obesos si estudien el consum proteic. Uns altres autors (Lissner 1989) no troben que els obesos infrareportin més que els no obesos, usant R24H. Ferrari (2002) mostra com els individus amb major IMC de l'estudi de calibratge EPIC tendeixen a infrareportar el seu consum energètic i els individus amb menor IMC a sobrereportar-lo. El percentatge d'individus identificats com a infrareportadors (aquells que consumeixen menys energia de la que el seu cos requereix segons la fórmula de Goldberg (1991)) a l'EPIC és per sota del 13% en tots els centres (excepte Grècia 20%) en homes i del 17% (Grècia 33%) en dones. El grau d'infrareport és heterogeni entre països però homogeni entre centres d'un mateix país. Cal dir, però, que aquestes estimacions s'han fet assumint que l'activitat física és constant per a tota la població, i potser els obesos són menys actius i necessiten menys energia. S'han proposat alguns models per corregir aquest biaix (Prentice 1996) basats en correccions a nivell individual que necessiten mesures repetides o l'ús de factors de calibratge específics per cada nivell d'IMC. Tot i així, no hi ha consens en el grau d'afectació sobre el factor de calibratge que provoca l'IMC (Ferrari 2002). En qualsevol cas no es pot provar aquesta assumpció en aquest projecte en no poder avaluar l'error de mesura d'*R*.

Altres assumpcions dels residus del model de calibratge: s'assumeix que no hi ha d'haver una relació dels residus del model de calibratge amb *T*. Si mirem la figura 21, en què es creuen els residus estudentitzats amb una estimació de *T*, els valors *R* del qüestionari de referència, no sembla que aquesta assumpció sigui acceptable. Però si tenim en compte que la variabilitat explicada pel model de calibratge és molt baixa, això vol dir que tot allò que no pot explicar el model de calibratge queda dipositat en el residu. D'aquí aquesta forta tendència dels residus envers el consum de referència.

Hem d'adonar-nos, però, que aquesta referència no és vàlida a nivell individual (d'aquí el mal ajust), però sí a nivell grupal (país). Per tant, hem de fixar-nos en què passa quan usem dades agregades (mitjanes) per país. Com es veu a la figura 22, tant per a homes com per a dones, no existeix cap tendència dels residus respecte el consum del qüestionari del R24H a nivell de país.

Figura 21. Diagrama de dispersió dels residus estudentitzats del model de calibratge de la carn i la carn mesurada usant el R24H. Homes espanyols.

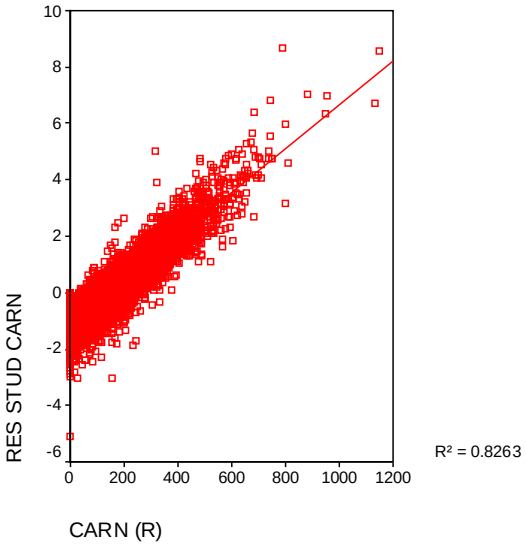
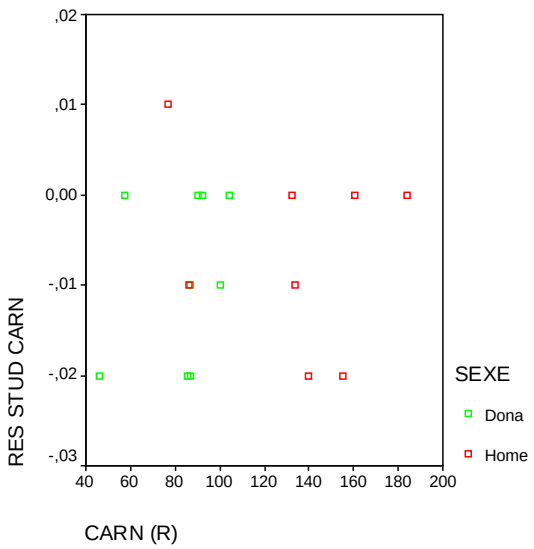


Figura 22. Diagrama de dispersió de les mitjanes per país dels residus estudentitzats del model de calibratge de la carn i la carn mesurada usant R24H.



Error no diferencial: donat que estem estudiant una cohort, en que tots els individus estan lliures de malaltia a l'inici de l'estudi (que és quan es mesura l'exposició) sembla difícil que pugui existir error diferencial en els qüestionaris *Q* i *R* (o sigui, que el grau d'error depengui del fet de si l'individu acabarà emmalaltint o no) (Rosner 1990). Tot i així, alguns símptomes i malalties precursors de CG poden fer variar la dieta i també la qualitat de les respostes (per exemple, se sap que els individus malalts tendeixen a recordar millor el que han menjat perquè els afecta més). L'exclusió dels individus diagnosticats en els dos primers anys donaria uns resultats en que es pot descartar definitivament la presència d'error diferencial. A més, la inclusió de covariables en el model de calibratge té en compte la influència que puguin tenir unes altres variables sobre l'error comès en mesurar l'exposició i la relació d'aquesta amb la malaltia (Rosner 1990). L'estimador que hem obtingut és major (més llunyà de la hipòtesi nul·la) quan excloem aquests individus diagnosticats prematurament.

Malaltia rara: podem definir com a malaltia rara aquella que té una prevalença baixa i és poc recurrent. En el cas de l'EPIC la prevalença de CG en el moment de fer aquest projecte era de 0,09%. En estudis de simulació, s'ha demostrat que $\hat{\beta}^*$ té poc biaix i

una cobertura de probabilitat apropiada si els OR no són majors de 3 en estudis de cohort i la malaltia d'estudi té una prevalença de fins al 5% (Rosner 1989).

Risc moderat: hem observat HR's al voltant de 2, que es poden considerar moderats.

Ingesta constant en el transcurs del temps: aquesta assumpció és impossible de verificar amb aquest estudi. Tot i així, és molt habitual en estudis de dieta en persones adultes.

Error de mesura baix: l'error de mesura del qüestionari de referència no és estimable com ja s'ha dit. El del qüestionari original es situa dins dels límits acceptables amb coeficients de calibratge entre 0,27 i 0,73 segons el país (habitualment s'accepten com a raonables coeficients entre 0,2 i 1).

La relació entre i ha de ser igual en la mostra de calibratge que en la resta de la cohort: com que la mostra de calibratge s'ha recollit de forma aleatòria entre tots els participants de la cohort, estratificant per edat i sexe, i tenint en compte l'estació de l'any i dia de la setmana, sembla probable que l'assumpció es compleixi.

$\hat{\beta}$ **i** $\hat{\lambda}$ **són independents:** aquesta assumpció seria sempre certa si uséssim mostres independents per estimar-les (Rosner 1990), o sigui, que la subpoblació de calibratge no pertanyés a la població de l'estudi principal. Això no s'ha fet així a l'EPIC però una forma de demostrar que aquesta assumpció és correcta és comparant les β^*'s obtingudes amb tota la cohort amb les obtingudes excloent de l'estudi principal a les persones que van participar a l'estudi de calibratge. S'esperaria que els resultats no variessin pràcticament. En aquest cas el HR varia d'1,97 (p=0,007) a 1,91 (p=0,013) al excloure la mostra de calibratge.

Mètodes de detecció dels individus malalts similars: diferències quant a completitud en la identificació de casos o l'elecció de la data de censura poden crear biaix a nivell intercohorts. En estratificar per país i ajustar per centre ens assegurem que diferències entre aquests no creen biaix a nivell intercohorts (Kaaks 1994a).

Pèrdues pel seguiment independents de l'exposició: aquesta hipòtesi no és demostrable per a aquest projecte. Tot i així és possible que algunes morts tinguin relació amb la dieta (altres càncers, malalties cardiovasculars,...). Repetint l'anàlisi excloent-ne els morts per causes diferents a les de CG obtenim un HR d'1,96 (p=0,007), pràcticament idèntic a l'observat en l'anàlisi amb tota la cohort, cosa que sembla descartar qualsevol efecte en el HR degut a pèrdues pel seguiment relacionades amb l'exposició.

Linealitat: la relació log-lineal entre el consum dietètic i el risc relatiu d'una malaltia és un model estàndard en anàlisis de cohort o cas-control (Kaaks 1995a), si bé també és comú l'ús de models lineals (Clayton 1988). En aquest estudi, hem observat com els residus de martingala són compatibles amb un model d'efecte lineal de la carn.

8.2 DISCUSSIÓ SOBRE EL MODEL USAT

En aquest estudi s'ha usat el R24H per calibrar. Kipnis (2002) adverteix, però, que l'ús de qüestionaris per calibrar no és el més apropiat. El model que proposa és el següent:

$$Q_{ij} = \phi_Q + \delta_Q T_i + q_i + \varepsilon_{ij} \text{ , } j=1,...,m_Q \text{ mesures repetides; } i=1,...,n \text{ individus} \quad (42)$$

$$R_{ij} = \phi_R + \delta_R T_i + r_i + u_{ij} \text{ , } j=1,...,m_R \text{ mesures repetides; } i=1, ..., n \text{ individus} \quad (43)$$

Si afegim un biomarcador, *L*, tindrem la següent equació addicional (Kipnis 2003):

$$L_{ij}=\phi_L+T_i+ v_{ij}, \text{ , } j=1,...,m_L \text{ mesures repetides; } i=1, ..., n \text{ individus} \quad (44)$$

Usant l'equació (44) podríem prescindir directament de (43) per calibrar *Q*, però necessitem (43) si volem saber quina és l'estructura de l'error d'*R* i la seva relació amb els errors de *Q (*Kipnis 2003*).* Kipnis (2002) també dóna una estimació del biaix amb que estimem λ al no usar biomarcadors com a referència:

$$\lambda_R = \lambda_L(\delta_R + \frac{1}{\delta_Q}\rho_{q,r}\sqrt{\frac{\sigma_q^2\sigma_r^2}{\sigma_L^4}}) \quad (45)$$

amb les mesures *L* considerades com *T*.

Com es veu, tant pel qüestionari original com pel de referència incorpora una estimació dels biaixos relacionats amb el consum real $\phi_Q+(\delta_Q-1)T_i$ i $\phi_R+(\delta_R-1)T_i$ respectivament per *Q* i *R*, q_i i r_i són els biaixos específics individuals i ε_{ij}, u_{ij} i v_{ij} són els errors aleatoris intraindividu. A més, el model de Kipnis permet que $\rho_{qr}>0$ i $\rho_{\varepsilon u}>0$. El model de Rosner seria un model particular d'aquest, assumint $\phi_R=0$, $\delta_R=1$ i $\sigma_r^2=0$. En defensa del model de Rosner cal veure que es compleix que $\text{cov}(\varepsilon_R,T)=0$ i $\text{cov}(\varepsilon_R, \varepsilon_Q)=0$ fins i tot si hi ha un cert biaix en *R* [$\phi_R\neq0$] (notis que encara cal però que $\delta_R=1$). Si *r* existeix (biaix a nivell individual en *R*), però *r* és independent de *q* tampoc es violen les anteriors assumpcions en el model de Rosner. Fins i tot es pot suportar que $\rho_{\varepsilon u}=0$ si els dos

mètodes s'han administrat no massa junts en el temps (Freedman 1990). La presència de biaix relacionat amb el consum real [$\delta_R \neq 1$] o la correlació entre els errors dels qüestionaris [$\rho_{qr} > 0$] invalidaria el mètode de referència pel model de Rosner. Com ja havíem vist, però, el model de Kipnis no és identificable sense una tercera mesura, no esbiaixada i amb errors no correlacionats amb els qüestionaris (biomarcador), i diverses mesures de *Q* i *R*. Cal dir que aquests biomarcadors han de tenir una relació coneguda amb la ingesta real, independent de la quantitat total de menjar ingerida i d'altres trets individuals com l'edat, el gènere, l'IMC,... (Kaaks 1995a). Per ara hi ha molt pocs marcadors que compleixin aquestes condicions. Fins i tot cal que el biomarcador es mesuri en un moment diferent al del qüestionari per evitar la correlació entre els errors aleatoris. Kipnis (2002) mostra com el seu mètode sempre troba coeficients de desatenuació menors que el mètode de Rosner que usem en aquest projecte (per tant el risc relatiu estimat pel model de Rosner en aquest projecte seria menor que el real, d'un 50% a un 200%), almenys així ho proven amb l'estudi d'ingesta de proteïnes, usant marcadors de nitrogen a l'orina. O sigui, que el mètode de Rosner que usem sobreestimaria la correlació entre *Q* i *T* (diria que *Q* és més bo del que en realitat és). Les pendents d'*R* amb *T* (δ_R) varien entre 0,34 i 0,77 a l'estudi de Kipnis (quan el model de Rosner les assumeix 1). La variància d'*r* (del biaix individual d'*R*) que Rosner assumeix 0 varia entre 0,01 i 0,05 en l'estudi de Kipnis, i la correlació entre els biaixos individuals *q* i *r* (que Rosner també assumeix 0) varia entre 0,35 i 0,95. No és clar si aquests resultats són extrapolables a uns altres nutrients o a proteïnes ajustades per energia (usant també un biomarcador per energia). Per tant, la correcció que Rosner efectua i que apliquem en aquest projecte pot ser incompleta (però en qualsevol cas millor que no fer-hi cap correcció).

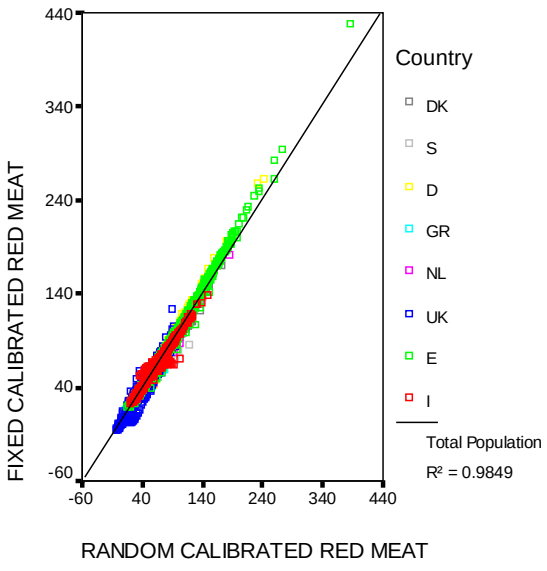
Finalment no hem aplicat cap transformació a les dades de consum de carn. Hem vist que l'ajust del model de calibratge no millorava especialment al aplicar la transformació arrel quadrada. La interpretació dels resultats serà sempre més fàcil si podem usar la variable original. A més cal tenir en compte que la mesura del R24H no és una estimació del consum real a nivell individual sinó a nivell grupal (o sigui E[R]=T, i no R=T). Això té més implicacions de les que pot semblar a primera vista. Per exemple, si fem la regressió de *R* amb *Q*, tenim E(R|Q). En el model de malaltia acabarem substituint la variable original *Q* per la variable calibrada E[R|Q][= E[T|Q], ja que

E[R]=T. Si apliquem l'arrel quadrada i fem la regressió de $\sqrt{R}$ amb $\sqrt{Q}$ tenim E[$\sqrt{R}$ | $\sqrt{Q}$], però, estem, implícitament, assumint que E[$\sqrt{R}$]=$\sqrt{T}$, o el que és el mateix, T= (E[$\sqrt{R}$])^2, que és diferent de l'assumpció bàsica E[R]=T (Boshuizen 2004).

En aquest projecte s'ha usat el model de calibratge de regressió lineal amb efectes fixos. En estudis anteriors, hem provat models multinivell (o sigui, d'efectes mixtos usant el centre com a factor aleatori). Els resultats per als HR obtinguts amb els dos models són força similars. A la figura 23 podem veure com les variables predites per un model de regressió lineal amb efectes fixos gairebé coincideixen amb un model d'efectes mixtos amb el centre com a factor aleatori.

L'ús de mètodes de calibratge més rudimentaris, com el calibratge additiu o multiplicatiu no permet tenir en compte variables de confusió (no permet corregir uns altres errors que no siguin els sistemàtics propis de cada centre o país). Per tant no semblen aconsellables, excepte com a possible solució per al calibratge de grups alimentaris amb un alt percentatge de no consumidors. Per últim, l'ús de mètodes de calibratge no lineals podria millorar la predicció de les variables calibrades, però com hem vist calen almenys dues mesures del R24H per poder aplicar aquests mètodes (Hoffmann 2002).

Figura 23. Diagrama de dispersió dels valors predits de carn vermella mitjançant models de regressió lineal amb efectes fixos i amb la variable centre com a factor aleatori (efectes mixtos). Cada color representa un país.



Hem vist com l'ajust del model de Cox és prou acceptable. Repetint els anàlisis amb models paramètrics (Weibull, log-normal) els resultats obtinguts són equivalents (no mostrat).

9. CONCLUSIÓ

Al llarg d'aquest document he presentat i aplicat el mètode del calibratge mitjançant una regressió lineal com a mètode d'estandardització i correcció dels biaixos dels HR obtinguts amb un model de Cox que relaciona el consum de carn i el CG en un estudi multicèntric.

Com hem vist és suficient emprar una sola mesura de referència en una submostra de la cohort per obtenir prediccions de consum per a tota la cohort, que un cop usades en el model de Cox ens donarà estimadors corregits del HR. Malgrat que el calibratge mitjançant l'ús d'un únic R24H com a mesura de referència pot incomplir alguna de les assumpcions bàsiques del calibratge (no correlació dels errors dels qüestionaris general i de referència, estimacions no esbiaixades del consum real a partir del qüestionari de referència), que no es poden provar amb les dades disponibles per aquest projecte, el calibratge serveix per corregir almenys una part del biaix amb què estimaríem els HR's si uséssim només les dades del qüestionari general.

És aconsellable l'ús de biomarcadors quan sigui possible, i l'estudi més acurat de mètodes alternatius que permetin emprar variables categòriques, un alt percentatge de no consumidors, relacions no lineals o presència d'errors correlacionats i biaix en els instruments de mesura.

Amb les dades disponibles hem vist com el consum de carn suposa un increment apreciable del risc de patir CG, independentment dels diferents tractaments a què hem sotmès les variables, i de com el calibratge allunyava el HR obtingut de la hipòtesi nul·la.

En definitiva, el calibratge és un recurs per minvar els efectes de l'error de mesura de la dieta en l'estimació dels paràmetres d'associació d'aquesta amb la malaltia.

10. REFERÈNCIES

Armstrong B (1985). Measurement error in generalized linear models. *Communications in Statistics, Series B* 14: 529-544.

Armstrong B, White E, Saracci R (1992). Principles of Measurement in Epidemiology. Oxford: Oxford Medical Publications, p. 63-64.

Berkson J (1950). Are there two regressions? *J Am Stat Assoc*, 45: 164-180.

Black AE, Coward WA, Cole TJ, Prentice AM (1996). Human energy expenditure in affluent societies: an analysis of 574 doubly-labelled water measurements. *Eur J Clin Nutr*, 50*:* 72–92.

Boshuizen HC, Ocké MC, Bueno de Mesquita HB (submitted 2004). Use of transformations in regression calibration: sense and nonsense. *Epidemiology.*

Buzzard M (1998). 24-hour dietary recall and food record methods. *Monogr Epidemiol Biostatist*, 30: 50-73.

Cameron ME, van Staveren WA (1988). Manual on Methodology for Food Consumption Studies. Oxford: Oxford University Press.

Carroll RJ, Stefanski LA (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *J Am Stat Assoc*, 85: 652-663.

Carroll RJ, Wand MP (1991). Semiparametric estimation in logistic measurement error models. *J Royal Stat Soc Series B,* 53: 573–585.

Carroll RJ, Freedman LS, Hartman AM (1996). Use of semiquantitative food frequency questionnaires to estimate the distribution of usual intake. *Am J Epidemiol*, 143(4): 392-404.

Carroll RJ, Freedman L, Pee D (1997). Design aspects of calibration studies in nutrition, with analysis of missing data in linear measurement error models. *Biometrics*, 53(4): 1440-1457.

Cisneros LM, Marquet L, Martí J, Pera JM, Saderra L (2001). Diccionari de la Qualitat. Barcelona: Associació d'Enginyers Industrials de Catalunya.

Clayton D (1988). Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In: Statistical models for longitudinal studies of health. Oxford: Oxford University Press, p. 301-331.

Cox DR (1972). Regression models and life-tables. *J Royal Stat Soc Series B*, 34: 187-220.

Ferrari P, Slimani N, Ciampi A, Trichopoulou A, Naska A, Lauria C, Veglia F, Bueno de Mesquita HB, Ocké MC, Brustad M, Braaten T, Tormo MJ, Amiano P, Mattisson I, Johansson G, Welch A, Davey G, Overvad K, Tjonneland A, Clavel-Chapelon F, Thiebaut A, Linseisen J, Boeing H, Hemon B, Riboli E (2002). Evaluation of under- and overreporting of energy intake in the 24-hour diet recalls in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Public Health Nutr*, 5(6B): 1329-1345.

Freedman LS, Schatzkin A, Wax J (1990). The impact of dietary measurement error on planning sample size required in a cohort study. *Am J Epidemiol*, 132: 1185–1195.

Freedman LS, Carroll RJ, Wax Y (1991). Estimating the relation between dietary intake obtained from a food frequency questionnaire and true average intake. *Am J Epidemiol*, 134(3): 310-320.

Freedman LS, Fainberg V, Kipnis V, Midthune D, Carroll RJ (2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*, 60: 172-181.

Friedenreich CM (1994). Methodologic issues for pooling dietary data. *Am J Clin Nutr*, 59(1 Suppl): 251S-252S.

Goldberg GR, Black AE, Jebb SA, Cole TJ, Murgatroyd PR, Coward WA, Prentice AM (1991). Critical evaluation of energy intake data using fundamental principles of energy physiology: 1. Derivation of cut-off limits to identify under-recording. *Eur J Clin Nutr*, 45(12): 569-581.

González CA, Pera G, Agudo A, Palli D, Krogh V, Vineis P, Tumino R, Panico S, Berglund G, Siman H, Nyrén O, Agren A, Martínez C, Dorronsoro M, Barricarte A, Tormo MJ, Quirós JR, Allen N, Bingham S, Day N, Miller A, Nagel G, Boeing H, Overvad K, Tjonneland A, Bueno de Mesquita HB, Boshuizen HC, Peeters P, Numans M, Clavel-Chapelon F, Helen I, Agapitos E, Lund E, Fahey M, Saracci R, Kaaks R, Riboli E (2003). Smoking and the risk of gastric cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Int J Cancer*, 107(4): 629-634.

Greenland S (1980). The effect of misclassification in the presence of covariates. *Am J Epidemiol*, 112(4): 564-569.

Greenland S (1987). Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev*, 9: 1-30.

Greenland S (1992). Divergent biases in ecologic and individual-level studies. *Stat Med*, 11(9): 1209-1223.

Haukka JK (1995). Correction for covariate measurement error in generalized linear models - a bootstrap approach. *Biometrics*, 51: 1127–1132.

Heitmann BL, Lessner L (1995). Dietary underreporting by obese individuals-is it specific or non-specific? BMJ, 311: 986-989.

Hoffmann K, Kroke A, Klipstein-Grobusch K, Boeing H (2002). Standardization of dietary intake measurements by nonlinear calibration using short-term reference data. *Am J Epidemiol*, 156(9): 862-870.

Huang Y, Wang CY (2001). Consistent functional methods for logistic regression with error in covariates. *J Am Stat Assoc,* 96 (456): 1469-1482.

Hunter D (1990). Biochemical indicators of dietary intake. In: Nutritional Epidemiology. New York: Oxford University Press, p. 143-216.

Jain MG, Rohan TE, Soskolne CL, Kreiger N (2003). Calibration of the dietary questionnaire for the Canadian Study of Diet, Lifestyle and Health cohort. *Public Health Nutr,* 6(1): 79-86.

Kaaks R, Plummer M, Riboli E, Esteve J, van Staveren W (1994). Adjustment for bias due to errors in exposure assessments in multicenter cohort studies on diet and cancer: a calibration approach. *Am J Clin Nutr,* 59(1 Suppl): 245S-250S.

Kaaks R, Riboli E, Esteve J, van Kappel AL, van Staveren WA (1994). Estimating the accuracy of dietary questionnaire assessments: validation in terms of structural equation models. *Stat Med,* 13(2): 127-142.

Kaaks R (1995). Calibration of dietary intake measurements in prospective cohort studies. *Am J Epidemiol,* 142(5): 548-556.

Kaaks R, Riboli E, van Staveren W (1995). Sample size requirements for calibration studies of dietary intake measurements in prospective cohort investigations. *Am J Epidemiol,* 142(5): 557-565.

Kaaks R (1997). Validation and calibration of dietary intake measurements in the EPIC project: methodological considerations. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol,* 26 Suppl 1: S15-S25.

Karvetti RL, Knuts LR (1985). Validity of the 24-hour dietary recall. *J Am Diet Assoc,* 85: 1437-1442.

Kipnis V (1999). Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. *Am J Epidemiol*, 150(6): 642-651.

Kipnis V, Midthune D, Freedman LS, Bingham S, Schatzkin A, Subar A, Carroll RJ (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications. *Am J Epidemiol*, 153(4): 394-403.

Kipnis V, Midthune D, Freedman L, Bingham S, Day NE, Riboli E, Ferrari P, Carroll RJ (2002). Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutr*, 5(6A): 915-923.

Kipnis V, Subar AF, Midthune D, Freedman LS, Ballard-Barbash R, Troiano RP, Bingham S, Schoeller DA, Schatzkin A, Carroll RJ (2003). Structure of dietary measurement error: results of the OPEN biomarker study. *Am J Epidemiol*, 158(1): 14-21.

Klesges RC, Eck LH, Ray JW (1995). Who underreports dietary intake in a dietary recall? Evidence from the Second National Health and Nutrition Examination Survey. *J Consult Clin Psychol*, 63(3): 438-444.

Korn EL, Graubard BI, Midthune D (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *Am J Epidemiol*, 145(1): 72-80.

Kuha J (1994). Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Statistics in Medicine*, 13: 1135-1148.

Kupper LL (1984). Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. *Am J Epidemiol*, 120(4): 643-648.

Kynast-Wolf G, Becker N, Kroke A, Brandstetter BR, Wahrendorf J, Boeing H (2002). Linear regression calibration: theoretical framework and empirical results in EPIC, Germany. *Ann Nutr Metab*, 46(1): 2-8.

Landin R, Freedman LS, Carroll RJ (1995). Adjusting for time trends when estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *Biometrics*, 51(1): 169-181.

Lissner L, Habicht JP, Strupp BJ, Levitsky DA, Haas JD, Roe DA (1989). Body composition and energy intake: do overweight women overeat and underreport? *Am J Clin Nutr*, 49(2): 320-325.

Madden JP, Goodman SJ, Guthrie HA (1976). Validity of the 24-hour recall. *J Am Diet Assoc*, 68(2): 143-147.

Morgan KJ, Johnson SR, Rizek RL, Reese R, Stampley GL (1987). Collection of food intake data: an evaluation of methods. *J Am Diet Assoc*, 87(7): 888-896.

Nyrén O, Adami HO (2002). Stomach cancer. In: Textbook of cancer Epidemiology. New York: Oxford University Press.

Peña D (2000). Diagnosis y validación del modelo de regresión múltiple. En: Estadística. Modelos y métodos. 2. Modelos lineales y series temporales. Madrid: Alianza Universidad Textos.

Piantadosi S, Byar D, Green S (1998). The ecological fallacy. *Am J Epidemiol*, 127: 893-904.

Plummer M (1994). Calibration in multi-centre cohort studies. *Int J Epidemiol*, 23(2): 419-426.

Plummer M, Kaaks R (2003). Commentary: An OPEN assessment of dietary measurement errors. *Int J Epidemiol*, 32(6): 1062-1063.

Prentice RL (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69: 331-342.

Prentice RL, Sheppard L (1989). Validity of international, time trend, and migrant studies of dietary factors and disease risk. *Prev Med*, 18(2): 167-179.

Prentice RL (1996). Measurement error and results from analytic epidemiology: dietary fat and breast cancer. *J Natl Cancer Inst*, 88(23): 1738-1747.

Rao CR (1973). Linear statistical inference and its applications. 2nd ed. New York: John Wiley and Sons.

Riboli E, Kaaks R (2000). Invited commentary: The challenge of multi-center cohort studies in the search for diet and cancer links. *Am J Epidemiol*, 151: 371-374.

Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, Charrondiere UR, Hemon B, Casagrande C, Vignat J, Overvad K, Tjonneland A, Clavel-Chapelon F, Thiebaut A, Wahrendorf J, Boeing H, Trichopoulos D, Trichopoulou A, Vineis P, Palli D, Bueno de Mesquita HB, Peeters PH, Lund E, Engeset D, González CA, Barricarte A, Berglund G, Hallmans G, Day NE, Key TJ, Kaaks R, Saracci R (2002). European Prospective Investigation into Cancer and Nutrition: Study populations and data collection . *Publ Health Nutr*, *5* (*6B*)*:* 1113–1124.

Richardson S, Gilks WR (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am J Epidemiol*, 138: 430–442.

Robins JM, Hsieh F, Newey W (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured data. *J Royal Stat Soc Series B,* 57: 409–424.

Rosner B, Willett WC (1988). Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. *Am J Epidemiol*, 127(2): 377-386.

Rosner B, Willett WC, Spiegelman D (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med*, 8(9): 1051-1069.

Rosner B, Spiegelman D, Willett WC (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error. *Am J Epidemiol*, 132: 734-745.

Rosner B, Gore R (2001). Measurement error correction in nutritional epidemiology based on individual foods, with application to the relation of diet to breast cancer. *Am J Epidemiol*, 154(9): 827-835.

SAS Institute (2001). The phreg procedure. In: SAS System Help. SAS release 8.02 [software]. Cary: SAS Institute Incorporation.

Schatzkin A, Kipnis V, Carroll RJ, Midthune D, Subar AF, Bingham S, Schoeller DA, Troiano RP, Freedman LS (2003). A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: results from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study. *Int J Epidemiol*, 32(6): 1054-1062.

Searle SR (1982). Matrix algebra useful for statistics. New York: John Wiley and Sons.

Slimani N, Deharveng G, Charrondiere RU, van Kappel AL, Ocké MC, Welch A, Lagiou A, van Liere M, Agudo A, Pala V, Brandstetter B, Andren C, Stripp C, van Staveren WA, Riboli E (1999). Structure of the standardized computerized 24-hour diet recall interview used as reference method in the 22 centers participating in the EPIC project. *Comput Methods Programs Biomed*, 58(3): 251-266.

Slimani N, Ferrari P, Ocké M, Welch A, Boeing H, Liere M, Pala V, Amiano P, Lagiou A, Mattisson I, Stripp C, Engeset D, Charrondiere R, Buzzard M, van Staveren W, Riboli E (2000). Standardization of the 24-hour diet recall calibration method used in the European Prospective Investigation into Cancer and Nutrition (EPIC): general concepts and preliminary results. *Eur J Clin Nutr*, 54(12): 900-917.

Slimani N, Kaaks R, Ferrari P, Casagrande C, Clavel-Chapelon F, Lotze G, Kroke A, Trichopoulos D, Trichopoulou A, Lauria C, Bellegotti M, Ocké MC, Peeters PH, Engeset D, Lund E, Agudo A, Larranaga N, Mattisson I, Andren C, Johansson I, Davey G, Welch AA, Overvad K, Tjonneland A, van Staveren WA, Saracci R, Riboli E (2002). European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study: rationale, design and population characteristics. *Public Health Nutr*, 5(6B): 1125-1145.

Slimani N, Bingham S, Runswick S, Ferrari P, Day NE, Welch AA, Key TJ, Miller AB, Boeing H, Sieri S, Veglia F, Palli D, Panico S, Tumino R, Bueno de Mesquita B, Ocké MC, Clavel-Chapelon F, Trichopoulou A, van Staveren WA, Riboli E (2003). Group level validation of protein intakes estimated by 24-hour diet recall and dietary questionnaires against 24-hour urinary nitrogen in the European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study. *Cancer Epidemiol Biomarkers Prev*, 12(8): 784-795.

Spector TD, Thompson SG (1991). The potential and limitations of meta-analysis. *J Epidemiol Community Health*, 45(2): 89-92.

Spiegelman D (1997). Measurement error correction for logistic regression models with an alloyed gold standard. *Am J Epidemiol*, 145: 184-196.

Spiegelman D, McDermott A, Rosner B (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr*, 65(4 Suppl): 1179S-1186S.

Spiegelman D (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Stat Med*, 20(1): 139-160.

Stata Corp. (2003). stcox-Fit Cox proportional hazards model. In: Stata Statistical Software: Release 8.0. College Station: Stata Corporation.

Stram DO, Hankin JH, Wilkens LR, Pike MC, Monroe KR, Park S, Henderson BE, Nomura AM, Earle ME, Nagamine FS, Kolonel LN (2000). Calibration of the dietary questionnaire for a multiethnic cohort in Hawaii and Los Angeles. *Am J Epidemiol*, 151(4): 358-370.

Stürmer T, Thurigen D, Spiegelman D, Blettner M, Brenner H (2002). The performance of methods for correcting measurement error in case-control studies. *Epidemiology*, 13(5): 507-516.

Therneau TM, Grambsch PM (2001). Residuals. In: Modeling Survival Data. Extending the Cox model. New York: Springer-Verlag.

Truett J, Cornfield J, Kannel W (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. *J Chron Dis*, 20: 511-524.

van Loon AJ, Goldbohm RA, van den Brandt PA (1998). Socioeconomic status and stomach cancer incidence in men: results from The Netherlands Cohort Study. *J Epidemiol Community Health,* 52: 166-171.

Wacholder S, Armstrong B, Hartge P (1993). Validation studies using an alloyed gold standard. *Am J Epidemiol*, 137(11): 1251-1258.

Ward MH, Sinha R, Heineman EF, Rothman N, Markin R, Weisenburger DD, Correa P, Zahm SH (1997). Risk of adenocarcinoma of the stomach and esophagus with meat cooking method and doneness preference. *Int J Cancer*, 71(1): 14-19.

White E (2003). Design and interpretation of studies of differential exposure measurement error. *Am J Epidemiol*, 157: 380-387.

Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, Hennekens CH, Speizer FE (1985). Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am J Epidemiol,* 122(1): 51-65.

Willett WC (1998). Nutritional epidemiology. 2nd ed. New York: Oxford University Press.

Witschi JC (1990). Short-term recall and recording methods. In: Nutritional epidemiology. New York: Oxford University Press, p. 53-68.

World Cancer Research Fund (1997). Stomach. In: Food, nutrition and the prevention of cancer: a global perspective. Washington: American Institute for Cancer Research.

Wynder EL, Hebert JR (1987). Homogeneity in nutritional exposure: an impediment in cancer epidemiology. *J Natl Cancer Inst,* 79(3): 605-607.

ANEXOS

A1. ÍNDEX D'ABREVIATURES I SÍMBOLS

a_i: taxa d'incidència de la malaltia a la cohort *i*.

c_i: nombre de casos a la cohort *i*.

corr: correlació.

cov: covariància.

e: error aleatori al regressar *T* amb *Q*.

e_Q: error total (sistemàtic i aleatori) de *Q*.

e_R: error total d'*R*. S'assumeix que coincideix amb ε_R.

exp: exponencial.

log: logaritme (si no s'especifica, logaritme neperià).

n_i: mida mostral de la cohort *i*.

p: p-valor.

p. ex.: per exemple.

q_i: biaix específic de l'individu *i* pel qüestionari general *Q*.

r_i: biaix específic de l'individu *i* pel qüestionari de referència *R*.

r^2: coeficient de determinació.

t: temps de seguiment.

u_{ij}: error aleatori del qüestionari de referència *R* per l'individu *i* rèplica *j*.

var: variància.

v_{ij}: error aleatori del biomarcador *L* per l'individu *i* rèplica *j*.

C: nombre total de casos entre *i* cohorts.

CG: càncer gàstric.

D: variable dicotòmica que expressa el fet de patir o no una malaltia.

E: esperança matemàtica.

EI/ER: raó energia consumida / energia requerida.

HD: història de dieta. S'usa com a *Q* en aquest projecte en alguns centres.

HR: *hazard ratio*. Raó de perills.

IC95%: interval de confiança al 95%.

IMC: índex de massa corporal.

I(t): taxa d'incidència en el temps *t*.

L: mesura aportada per un biomarcador.

N: nombre d'individus a l'estudi de calibratge.

N_i: nombre de rèpliques d'*R* de l'individu *i* de l'estudi de calibratge.

OR: *odds ratio*. Raó d'avantatges.

Pr: probabilitat.

P99: percentil 99.

Q: variable contínua que expressa el consum estimat habitual d'un aliment o nutrient. Es mesura amb un qüestionari general aplicat a tota la mostra, subjecte a errors de mesura.

QFA: qüestionari de freqüència alimentària. S'usa com a *Q* en aquest projecte per la majoria de centres.

Q1: primer quartil de *Q* calculat usant tota la cohort, específic per gènere.

Q2: segon quartil de *Q* calculat usant tota la cohort, específic per gènere.

Q3: tercer quartil de *Q* calculat usant tota la cohort, específic per gènere.

Q4: quart quartil de *Q* calculat usant tota la cohort, específic per gènere.

R: variable contínua que expressa el consum estimat d'un aliment o nutrient. Es mesura amb un qüestionari de referència, que se suposa que permet obtenir mesures no esbiaixades del consum real (T), almenys a nivell grupal.

RR: risc relatiu.

R24H: record de 24 hores. S'usa com a *R* en aquest projecte.

SD: desviació tipus.

SE: error estàndard.

T: variable contínua que expressa el consum real habitual d'un aliment o nutrient.

U: vector de variables d'ajust mesurades sense error.

X: E[T|Q]. Consum real mig donat el consum observat *Q*. Consum calibrat (o predit pel model de calibratge).

X1: primer quartil d'*X* calculat usant tota la cohort, específic per gènere.

X2: segon quartil d'*X* calculat usant tota la cohort, específic per gènere.

X3: tercer quartil d'*X* calculat usant tota la cohort, específic per gènere.

X4: quart quartil d'*X* calculat usant tota la cohort, específic per gènere.

α: terme constant en la relació entre malaltia i dieta mesurada amb error.

α^*: terme constant corregit en la relació entre malaltia i dieta real.

α': terme constant al regressar la dieta real *T* amb la dieta mesurada pel qüestionari *Q*.

β: associació entre malaltia i dieta mesurada amb error.

β^*: associació corregida entre malaltia i dieta real.

$\hat{\beta}_B$: estimador a nivell ecològic de la relació entre dieta real i malaltia.

$\hat{\beta}_O$: estimador global de la relació entre dieta i malaltia que té en compte la relació a nivell individual i a nivell ecològic.

$\hat{\beta}_W$: efecte ponderat de la dieta mesurada amb error sobre una malaltia entre diverses cohorts (ponderació dels efectes intracohort).

β_1 : associació entre malaltia i dieta mesurada amb error (quan es separen els efectes de la dieta i d'altres covariables).

β_1^*: associació corregida entre malaltia i dieta real (quan es separen els efectes de la dieta i d'altres covariables).

β_2 : associació entre malaltia i altres variables d'ajust diferents de la dieta (quan es separen els efectes de la dieta i d'altres covariables) quan la dieta està mesurada amb error.

β_2^*: associació corregida entre malaltia i altres variables d'ajust diferents de la dieta (quan es separen els efectes de la dieta i d'altres covariables).

δ_Q: pendent al regressar *Q* amb *T*. Error sistemàtic proporcional de *Q*.

δ_R: pendent al regressar *R* amb *T*. Error sistemàtic proporcional d'*R*.

ε_{ij} :error aleatori del qüestionari de mesura *Q* per l'individu *i* rèplica *j*.

ε_Q :error aleatori del qüestionari de mesura *Q*.

ε_R: error aleatori del qüestionari de referència *R*.

ϕ_Q: terme constant al regressar *Q* amb *T*. Error sistemàtic constant de *Q*.

ϕ_R: terme constant al regressar *R* amb *T*. Error sistemàtic constant d'*R*.

λ: coeficient de calibratge o de desatenuació. Associació entre *T* (o *R*) i *Q*.

λ_1: Associació entre la dieta real i la dieta mesurada pel qüestionari general (quan es separen els efectes de la dieta i d'altres covariables).

λ_2: Associació entre la dieta real i altres variables d'ajust diferents de la dieta (quan es separen els efectes de la dieta i d'altres covariables).

μ_i : mitjana del consum a la cohort *i*.

μ_Q :mitjana del consum mesurat amb el qüestionari general.

μ_T :mitjana del consum real.

ρ_{qr} : correlació entre els errors específics individuals de *Q* i *R*.

ρ_{QR} : correlació entre Q i R.

ρ_{QT} : correlació entre Q i T.

$\rho_{\varepsilon u}$: correlació entre els errors aleatoris de Q i R.

σ_q : desviació estàndard dels errors específics individuals de Q.

σ_r : desviació estàndard dels errors específics individuals d'R.

σ_L : desviació estàndard d'L.

σ_Q : desviació estàndard de Q.

σ^2_T : variància del consum real.

$\sigma^2_{\varepsilon Q}$: variància d'ε_Q .

τ_i : desviació estàndard del consum a la cohort i.

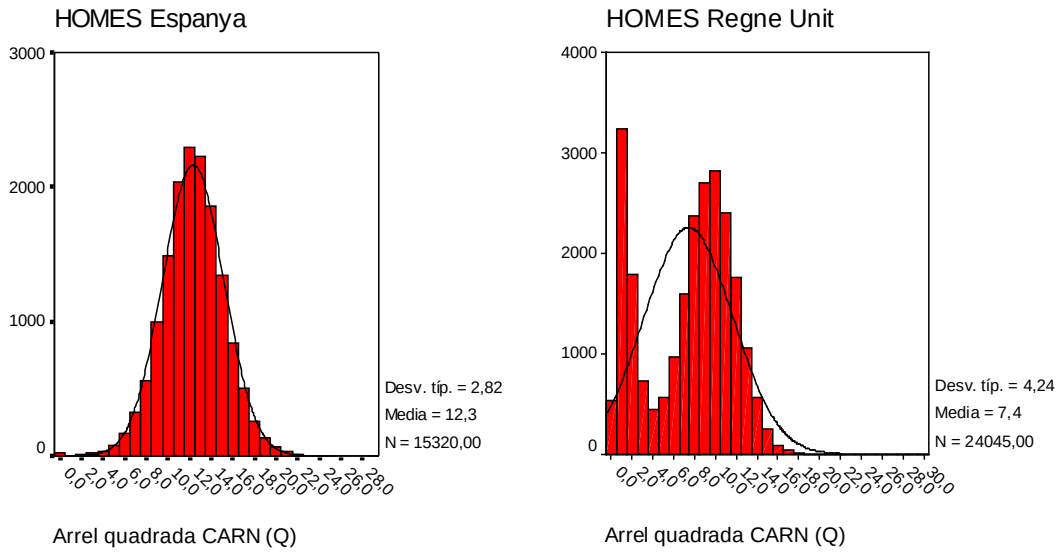
A2. MODIFICACIONS AL MODEL DE CALIBRATGE ORIGINAL

A2.1 TRANSFORMACIONS DE LES VARIABLES

Una de les possibles correccions al procés de calibratge seria l'ús de transformacions en les variables. Si mirem quina és la transformació de Box-Cox que ens pot normalitzar i fer més constant la variància de les dades obtenim valors de 0,45 i 0,33 per la carn del R24H i de 0,65 i 0,64 per la carn del QFA/HD per homes i dones respectivament.

Provem de transformar la variable carn fent l'arrel quadrada, tant pel R24H com pel QFA/HD i observem (figura A1) el comportament d'aquesta variable transformada per als homes d'Espanya (que coincideix amb el de la resta de centres i per a dones, excepte al Regne Unit). Podem veure com l'arrel quadrada de la variable QFA/HD sembla compatible amb una distribució normal, excepte per al Regne Unit.

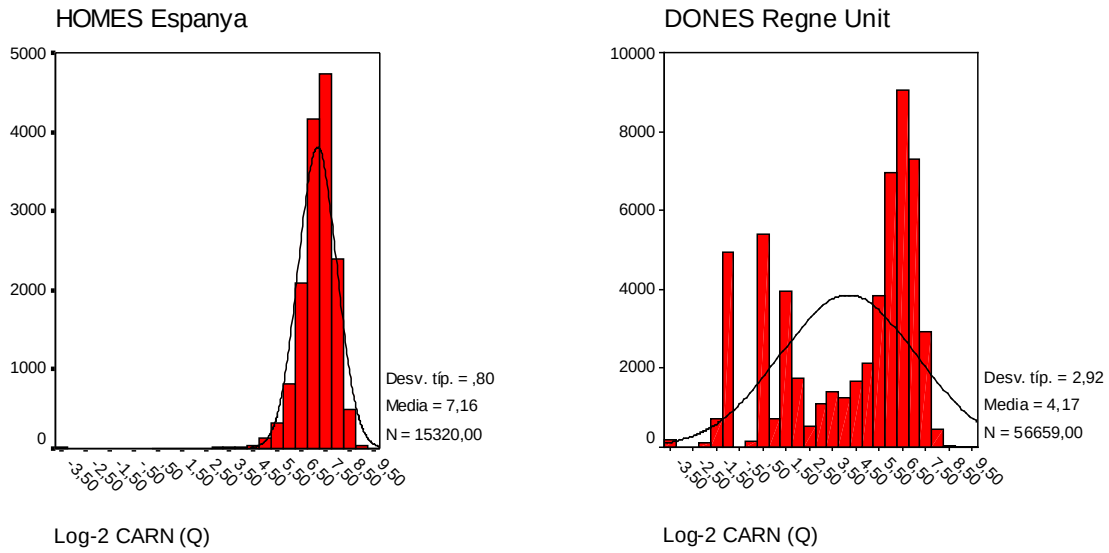
Figura A1. Histogrames de l'arrel quadrada del consum de carn QFA/HD en homes d'Espanya i Regne Unit.



Una transformació habitual de les dades de consum d'aliments és la logarítmica, tot i que en aquest cas la transformació de Box-Cox estimada no aconsella aquesta

transformació. A tall d'exemple mostrem l'efecte que té aquesta transformació logarítmica (en base 2) a Espanya, que té el mateix comportament que la resta de països, excepte les dones del Regne Unit. Com es veu, la cua que teníem a la dreta es manté ara a l'esquerra (figura A2).

Figura A2. Histogrames del logaritme en base 2 del consum de carn QFA/HD en homes d'Espanya i dones del Regne Unit.



Si mirem l'efecte de la transformació arrel quadrada a les dades R24H observem l'important efecte dels no consumidors (la transformació no pot corregir aquest efecte ja observat a la variable original). Veiem la gràfica per a Espanya, que és semblant a la resta de països. La transformació logarítmica en qualsevol cas no millora la variable no transformada ni l'arrel quadrada (figura A3).

En creuar la variable del QFA/HD amb la del R24H no s'aprecia cap millora important en els coeficients r^2, tant per mitjà de la transformació amb arrel quadrada com logarítmica. De nou, s'observa el comportament especial dels no consumidors del R24H. Es mostren els resultats per a Espanya (figura A4). No hi ha millores importants respecte a l'ús de les variables no transformades per a cap país.

Figura A3. Histogrames de l'arrel quadrada i logaritme en base 2 del consum de carn R24H en homes d'Espanya.

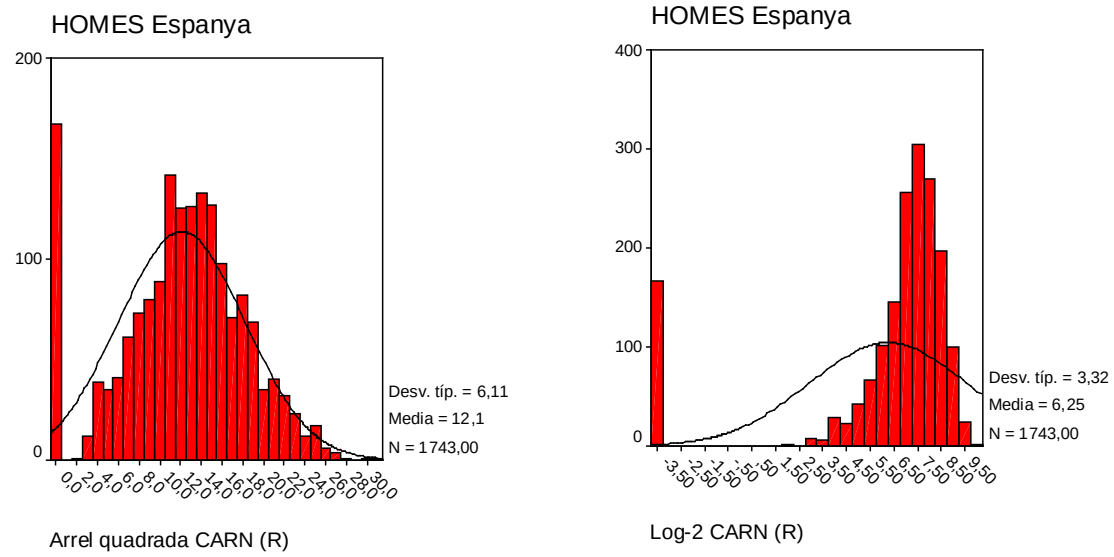
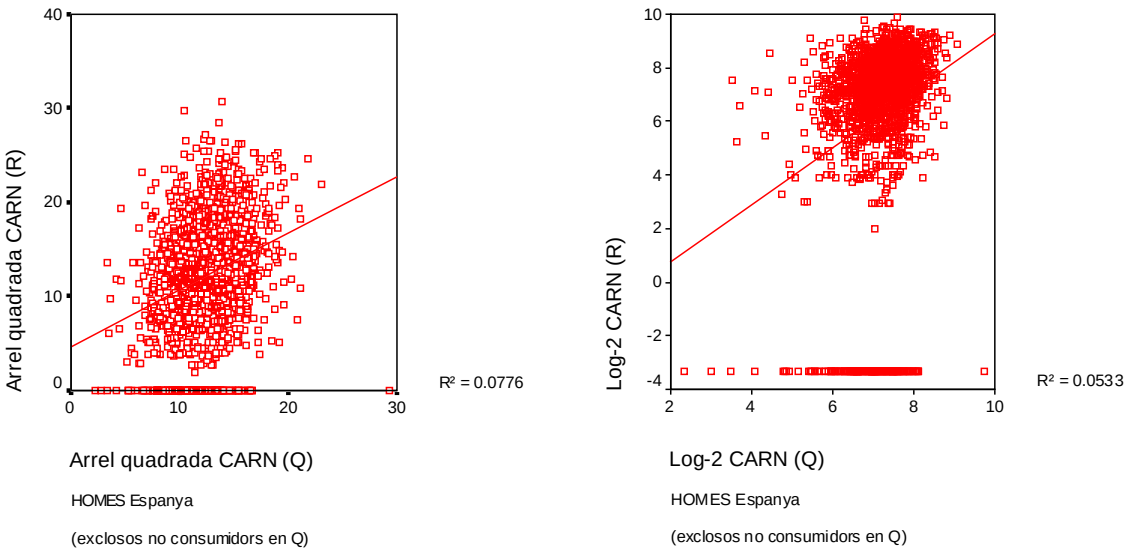


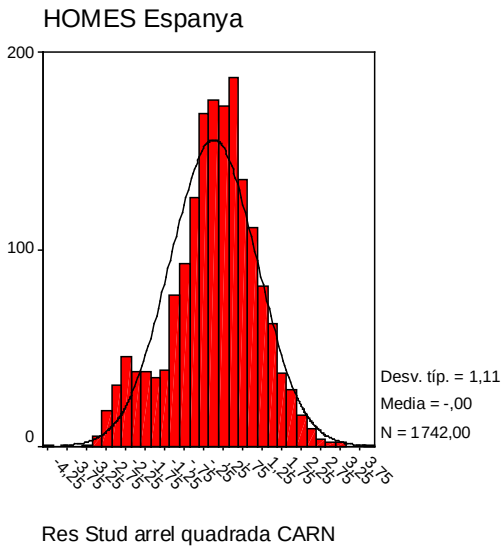
Figura A4. Diagrames de dispersió entre el consum de carn mesurat amb el R24H (R) i el QFA/HD (Q) per als homes d'Espanya aplicant les transformacions per arrel quadrada i per logaritme en base 2.

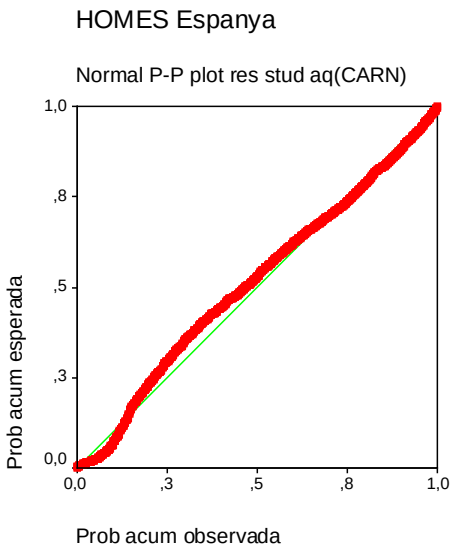


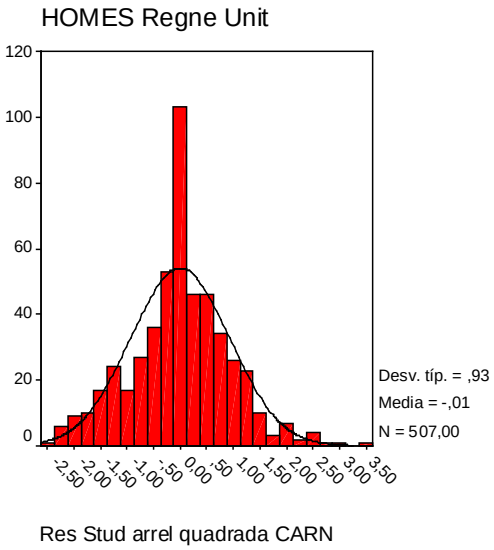
Si mirem els residus observem una lleugera millora de la normalitat en aplicar l'arrel quadrada, ja sigui mitjançant els histogrames o els dibuixos de probabilitat normal (es

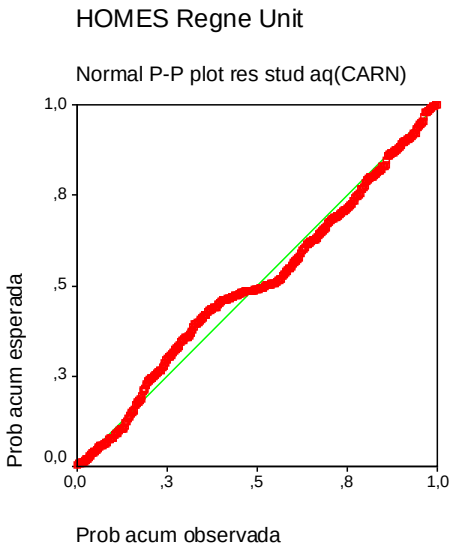
mostra el gràfic per a Espanya, similar al de la resta de centres, excepte el Regne Unit i Grècia, on tot i la transformació les dades són lluny de la normalitat) (figura A5). Per país, els gràfics són molt semblants per gènere.

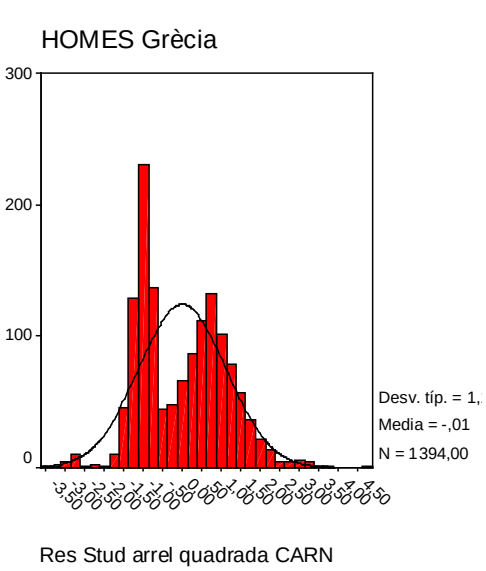
Figura A5. Histogrames i dibuixos de probabilitat normal dels residus del calibratge de carn usant la transformació per arrel quadrada per als homes d'Espanya, el Regne Unit i Grècia.

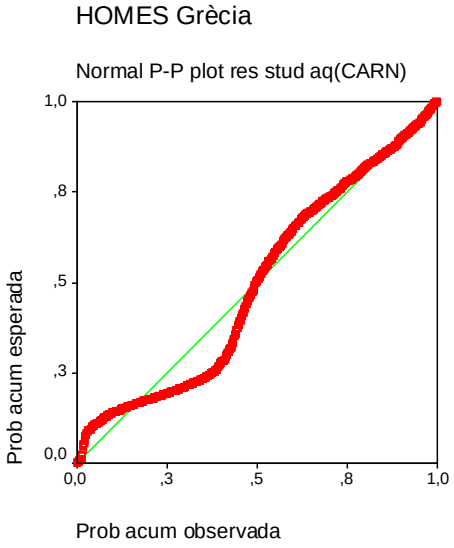






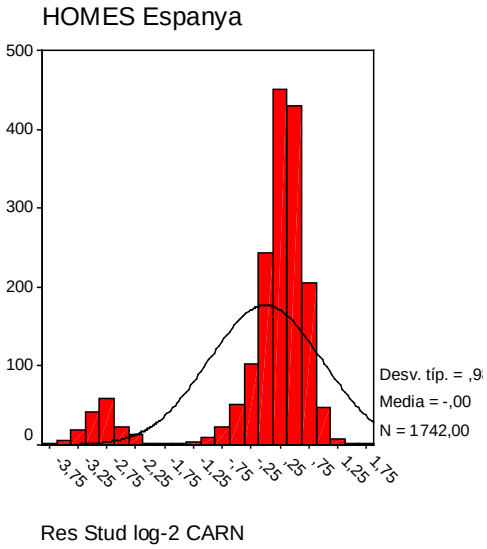


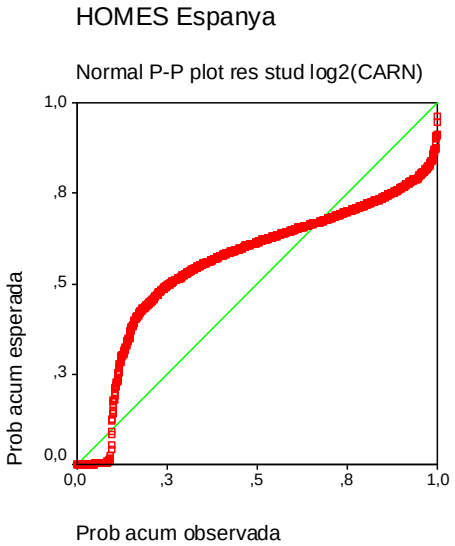


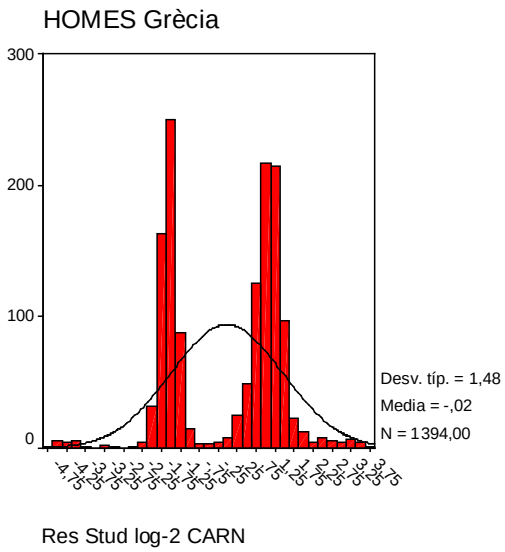


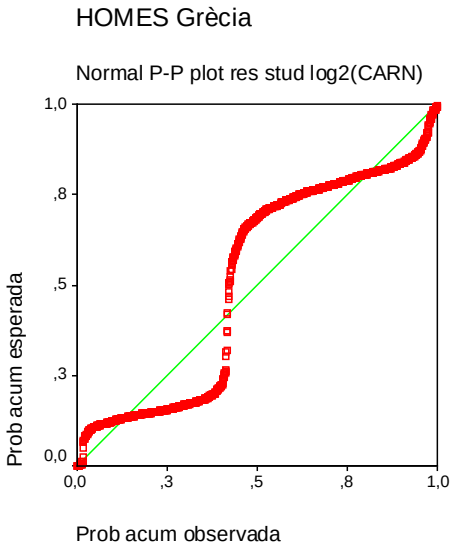
La situació no millora en absolut aplicant la transformació logarítmica. Lluny de normalitzar-se, els residus presenten una distribució bimodal, corresponent als no consumidors (part esquerra de l'histograma) i consumidors (part dreta) de carn en el R24H. Es mostren els gràfics per a Espanya, compatibles amb la resta de països excepte Grècia (figura A6). Els gràfics de cada país són molt similars per a homes i dones.

Figura A6. Histogrames i dibuixos de probabilitat normal dels residus del calibratge de carn usant la transformació logarítmica en base 2 per als homes d'Espanya i Grècia.



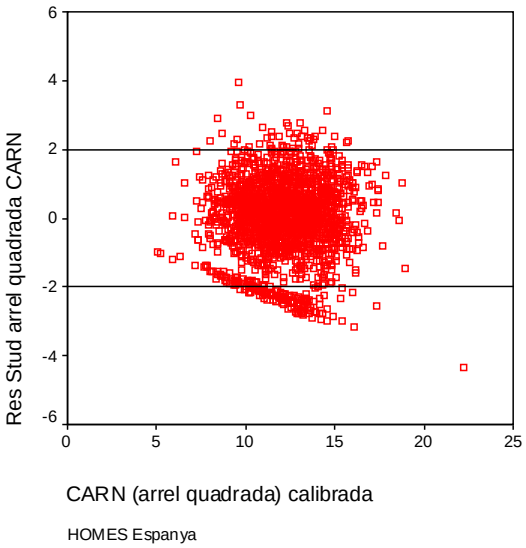


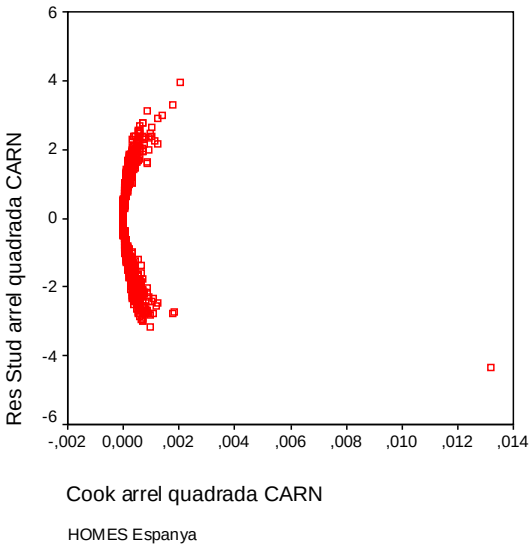


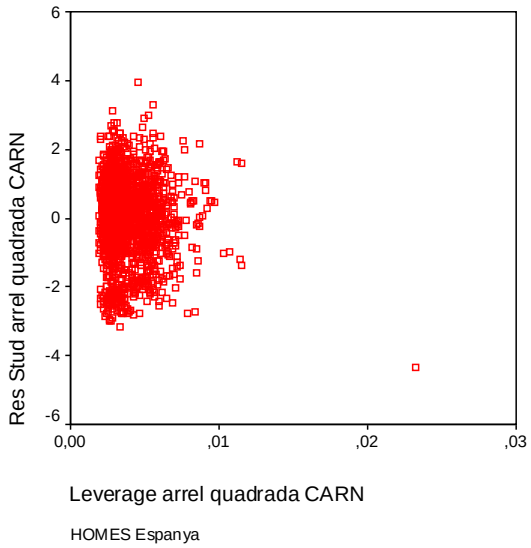


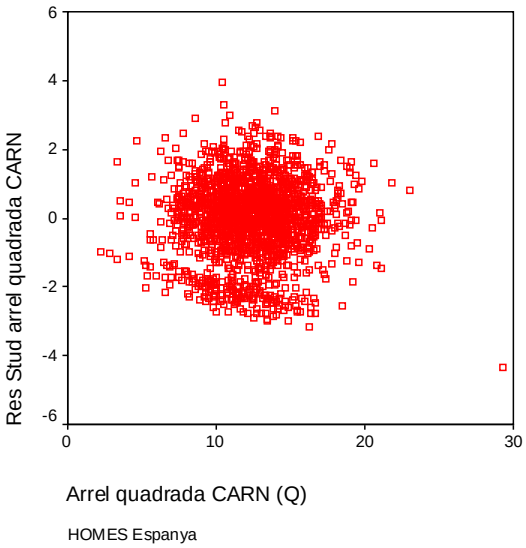
No observem una millora important creuant els residus estudentitzats amb la variable predita (calibrada), usant la transformació per arrel quadrada. L'efecte dels zeros (els punts que formen una recta a sota del núvol) persisteix i continua havent-hi una certa preponderància de residus positius. De nou s'observa la presència d'*outliers* (uns quants menys, pel fet que l'arrel quadrada comprimeix les dades) i punts influents. Es mostren els resultats per a Espanya, compatibles amb la resta de països (figura A7).

Figura A7. Diagrames de dispersió dels residus estudentitzats del calibratge de carn usant la transformació per arrel quadrada respecte a la variable predita (calibrada), coeficient de Cook, *leverage* i consum de carn QFA/HD per als homes d'Espanya.



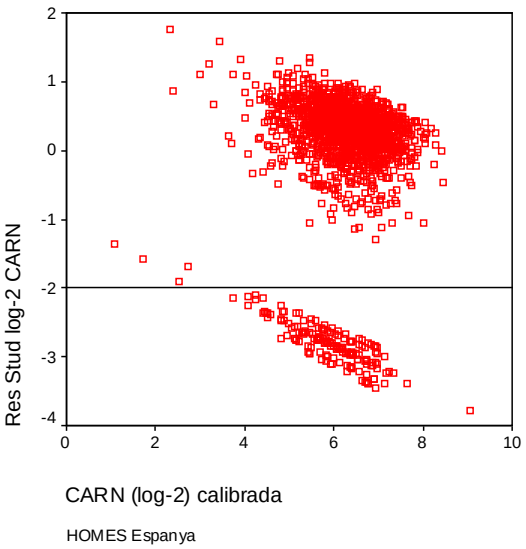


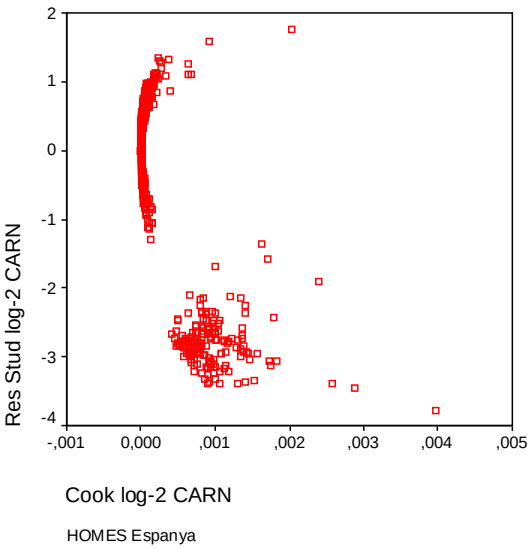


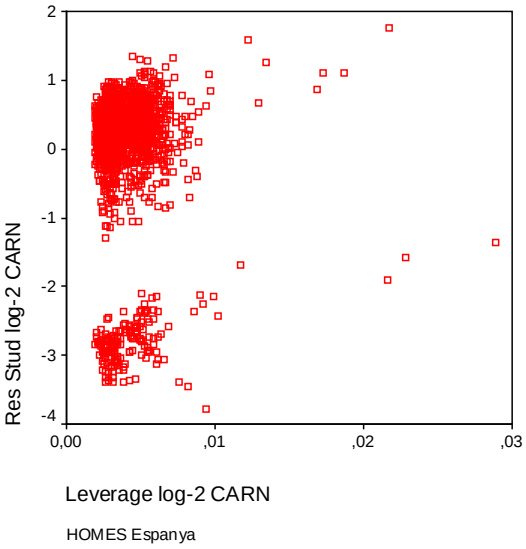


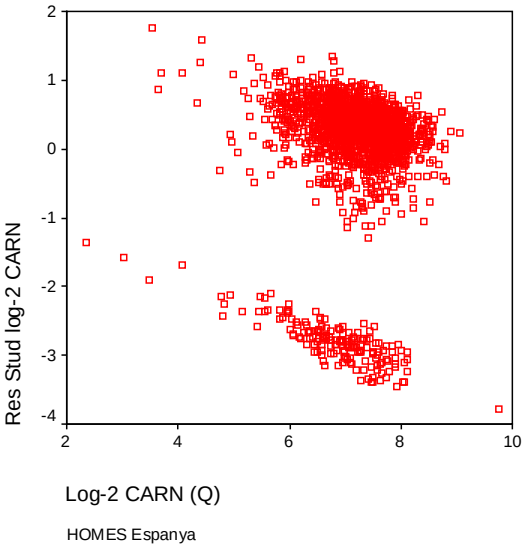
En el cas de la transformació logarítmica es veu clarament la diferència entre els no consumidors del R24H i els consumidors i no s'observa cap millora respecte al model no transformat. Es mostren les gràfiques per als homes espanyols, comparables a les de la resta de països i a les de les dones (figura A8).

Figura A8. Diagrames de dispersió dels residus estudentitzats del calibratge de carn usant la transformació logarítmica en base 2 respecte a la variable predita (calibrada), coeficient de Cook, *leverage* i consum de carn QFA/HD per als homes d'Espanya.









Un cop explorades les distribucions de la variable carn (tant del QFA/HD com del R24H), tant usant la variable original, com la transformació per arrel quadrada o logarítmica, i també el comportament de la variable predita i els seus residus, *outliers* i punts influents, no trobem una millora significativa pel fet d'aplicar una transformació. La interpretació dels resultats serà sempre més fàcil si podem usar la variable original.

A2.2 EXCLUSIÓ D'*OUTLIERS*

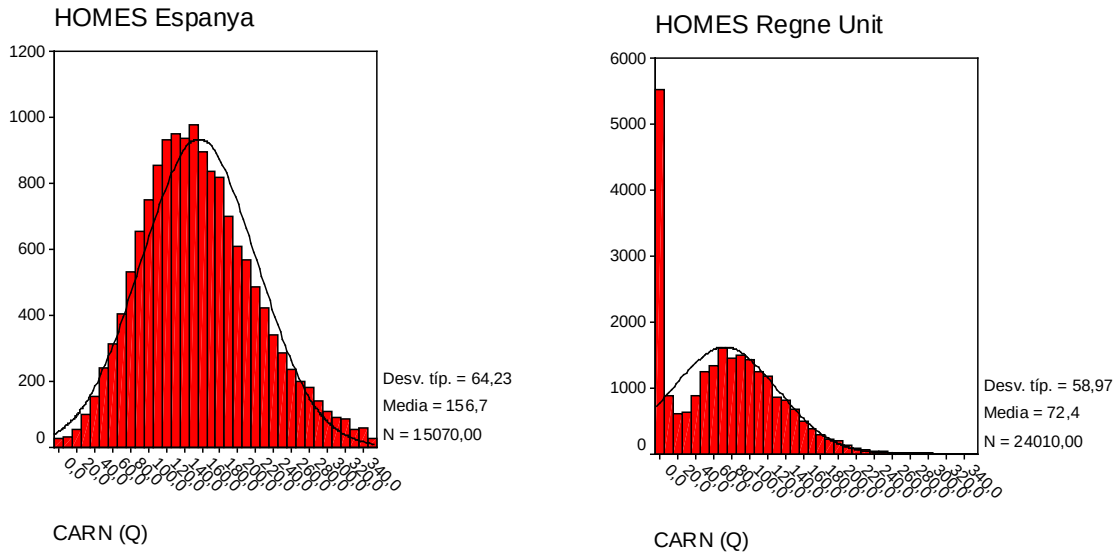
En una mostra grossa, com la que tenim en aquest estudi, és poc probable que uns quants valors extrems modifiquin la relació entre variables. S'ha passat un primer filtre en excloure els percentils extrems de la raó EI/ER. Tot i així encara queden valors grans (no hi poden haver *outliers* per l'esquerra perquè el consum mínim és 0) que podrien influir en els resultats. Una forma ràpida de veure l'efecte d'aquests valors és repetir el calibratge sense els *outliers.*

En aquest cas repetirem l'anàlisi per als homes i dones, excloent els que consumeixen més de 350 g/dia i 250 g/dia de carn en QFA/HD o més de 550 g/dia i 350g/dia en R24H respectivament per homes i dones (que equival al percentil 99 en cada cas aproximadament). El valor de R24H és més alt ja que es refereix al consum puntual d'un dia. La variació en els coeficients de desatenuació respecte a l'ús de les variables

sense exclusions és baixa. El país que més varia és Grècia (homes) que passa de 0,56 a 0,49.

Si ens fixem en els histogrames de la carn corresponent al QFA/HD observem que la distribució és més simètrica i normal un cop exclosos els grans consumidors que abans d'excloure'ls. Això passa a tots els països (es mostra Espanya) excepte al Regne Unit (molts no consumidors i cua per la dreta) (figura A9). Per les dones passa el mateix, però tallant el consum en 250 grams/dia.

Figura A9. Histogrames del consum de carn QFA/HD per homes d'Espanya i el Regne Unit després d'excloure els consumidors de més de 350 g/dia.



Si mirem els histogrames de carn reportada en els R24H s'observa que les cues per la dreta que teníem sense excloure els grans consumidors s'escurcen, però el pes dels no consumidors continua essent molt important. Aquest patró s'observa aproximadament a tots els països (es mostra Espanya) , excepte a Grècia i el Regne Unit (tenen el mateix patró) en què els no consumidors encara tenen més pes (figura A10). Per a les dones les gràfiques són similars a la dels homes, però limitades a 350 grams/dia.

En creuar les 2 mesures excloent-ne els grans consumidors no hi ha cap canvi evident en el núvol de punts ni en l'ajust de rectes de regressió si ho comparem amb els gràfics de dispersió sense excloure els grans consumidors (es mostra Espanya, però tots els

països i les dones tenen un comportament molt similar al que tenien sense excloure individus) (figura A11).

Figura A10. Histogrames del consum de carn R24H per homes d'Espanya i el Regne Unit després d'excloure els consumidors de més de 550 g/dia.

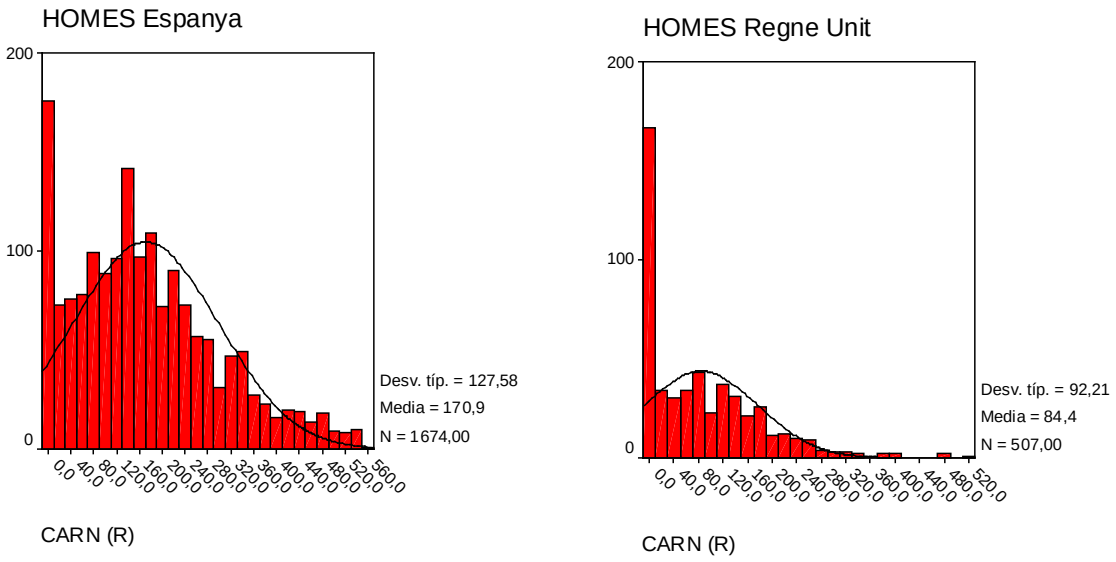
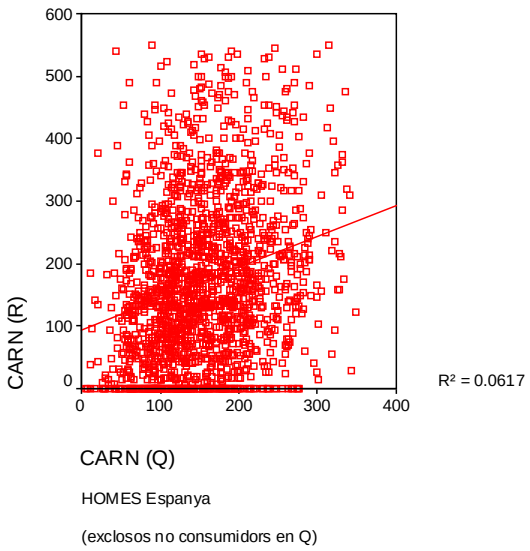
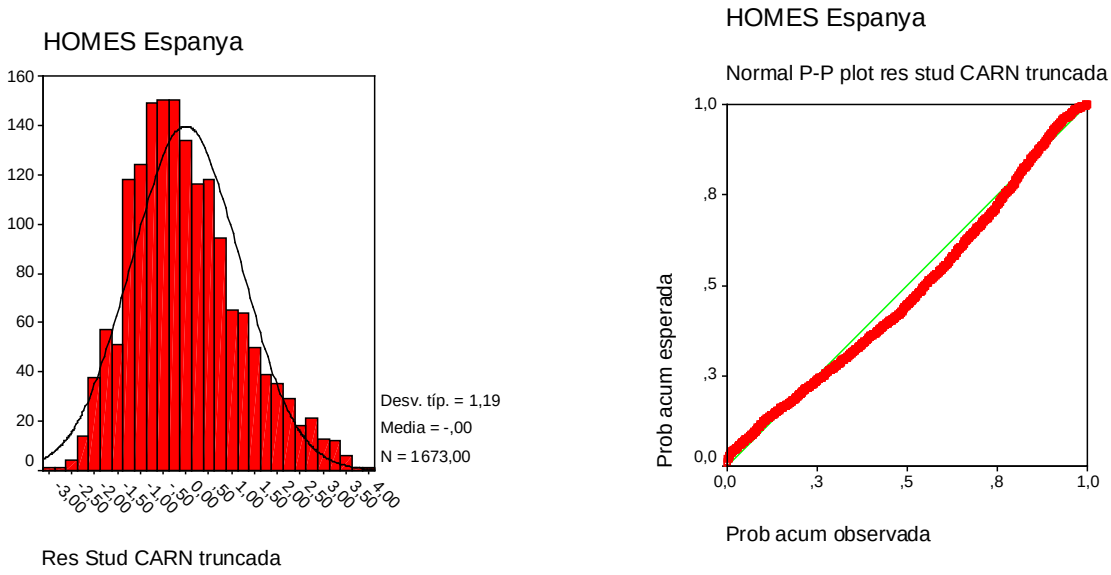


Figura A11. Diagrama de dispersió entre el consum de carn R24H (R) i el consum QFA/HD (Q) per a homes d'Espanya després d'excloure'n els consumidors de més de 350 g/dia en el QFA/HD i 550 g/dia en el R24H.



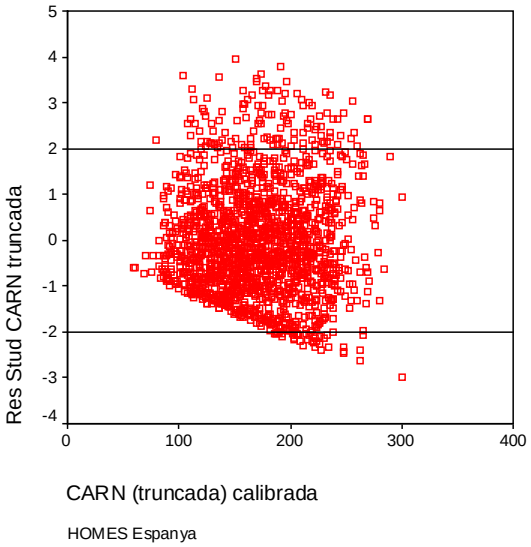
Els residus de la variable truncada s'assemblen als de l'original, però tenen un rang inferior (p. ex. per Espanya [homes] el rang passa de 12 a 7) (figura A12). El comportament és similar a tots els països, per a homes i dones, excepte el Regne Unit i Grècia. En aquests centres, però, el comportament dels residus també és molt similar a l'observat amb la variable sense excloure grans consumidors (figures 11 i 12).

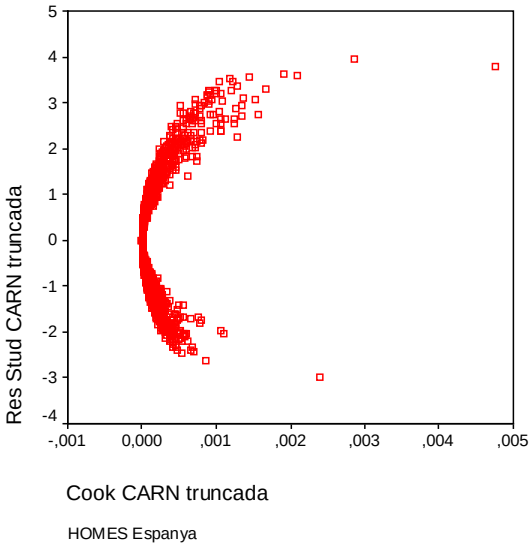
Figura A12. Histograma i dibuix de probabilitat normal dels residus del calibratge de carn després d'excloure'n els consumidors de més de 350 g/dia en el QFA/HD i 550 g/dia en el R24H per als homes d'Espanya.

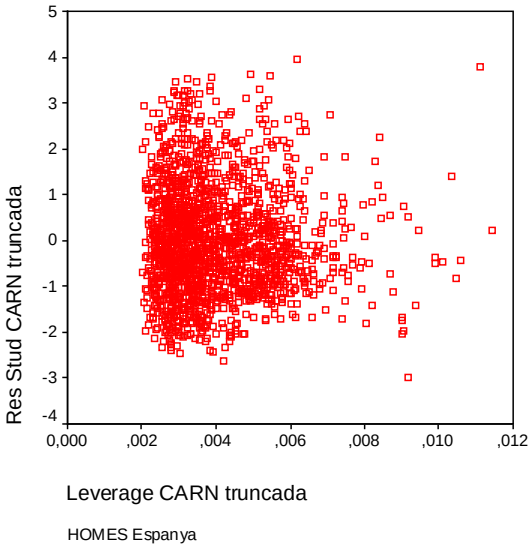


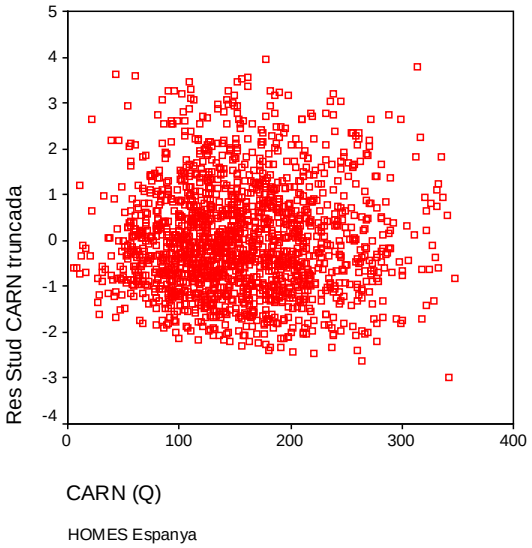
El *plot* de residus estudentitzats respecte a la variable calibrada té un comportament gairebé idèntic al que tenia abans d'excloure'n els grans consumidors, per a tots els països. Només desapareixen la majoria dels *outliers*. Si mirem els gràfics de dispersió dels residus estudentitzats respecte als coeficients de Cook, *leverage* i carn QFA/HD no s'hi aprecien gairebé canvis en eliminar els grans consumidors, excepte que els valors màxims de Cook i *leverage* disminueixen bastant en general (per tant, hi ha molts menys valors influents a *priori* i a *posteriori*). Es mostren els gràfics corresponents als homes d'Espanya, equiparables a la resta de països i als de les dones (figura A13).

Figura A13. Diagrames de dispersió dels residus estudentitzats del calibratge de carn respecte a la variable predita (calibrada), coeficient de Cook, *leverage* i consum de carn QFA/HD després d'excloure els consumidors de més de 350 g/dia en el QFA/HD i 550 g/dia en el R24H per als homes d'Espanya.









A2.3 RECODIFICACIÓ DELS ZEROS DEL QFA/HD

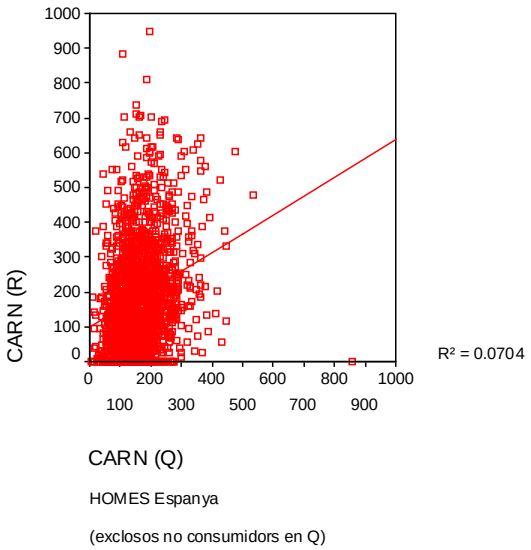
La definició de "no consumidor" d'un aliment no és senzilla. Fins ara hem considerat aquells que reporten una ingesta habitual (qüestionari QFA/HD) de 0 grams al dia. Però

poden aparèixer valors molt petits de consum provinents de receptes (habitualment receptes estàndard, que l'enquestat no podia modificar (en alguns centres això sí que era possible). Per exemple, un vegetarià que mengi una pizza podria estar ingerint carn (o se li podria imputar carn si la pizza té aquest ingredient per defecte, tot i que l'enquestat potser menja pizza sense carn). Sigui com sigui, la quantitat de carn provinent d'una pizza deu ser relativament petita. Podem intentar reclassificar els consumidors de quantitats petites com a no consumidors i observar si hi ha canvis. Arbitràriament, els investigadors de l'EPIC han triat consumidors habituals de menys de 3 grams al dia de carn com a "no consumidors".

Els coeficients de desatenuació gairebé no varien en excloure els consumidors de menys de 3 g/dia. El Regne Unit (que és on hi ha menys consumidors de carn) és l'únic país on hi ha un cert canvi (de 0,51 a 0,46 en homes i de 0,33 a 0,26 en dones).

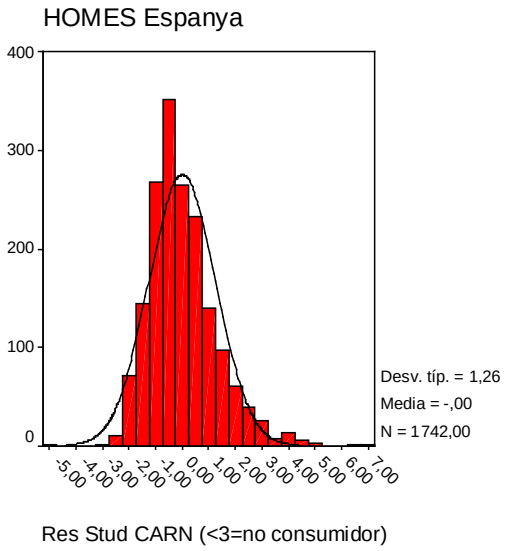
L'ajust del model no varia gaire en excloure els baixos consumidors. Només el Regne Unit pateix algun canvi (el r^2 baixa de 0,18 a 0,07 en homes). Es mostra només el gràfic per als homes d'Espanya (figura A14).

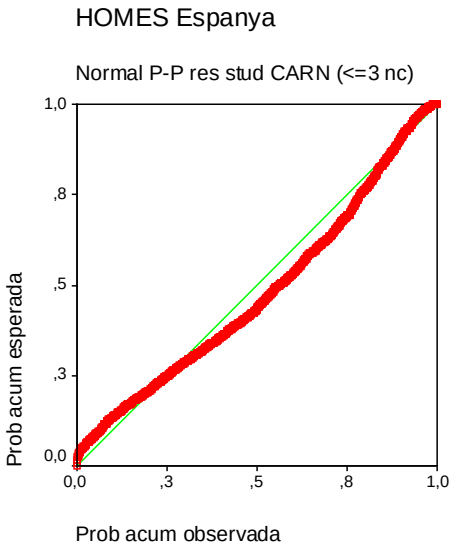
Figura A14. Diagrama de dispersió entre el consum de carn R24H (R) i el consum QFA/HD (Q) en homes d'Espanya després d'excloure els consumidors de menys de 3 g/dia en el QFA/HD.

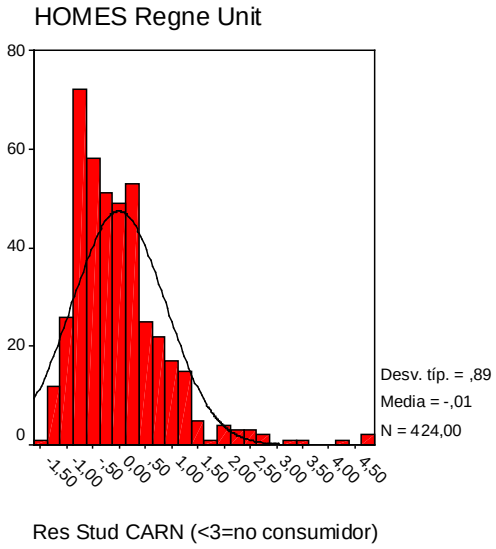


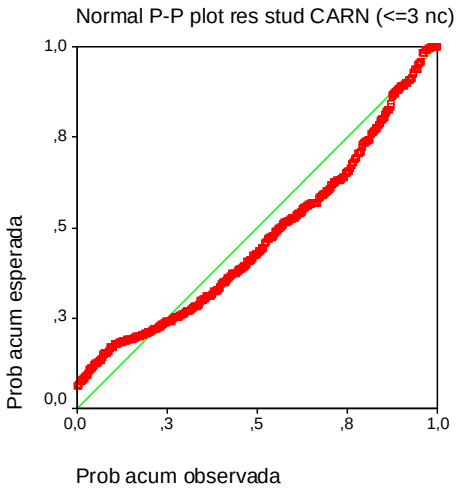
Els residus es comporten de la mateixa forma que es comportaven abans d'excloure els baixos consumidors, com es pot veure al gràfic per als homes d'Espanya. Només s'aprecia algun canvi en la distribució dels residus del Regne Unit (figures A15 i A16).

Figura A15. Histogrames i dibuixos de probabilitat normal dels residus del calibratge de carn després d'excloure els consumidors de menys de 3 g/dia en el QFA/HD per als homes d'Espanya i el Regne Unit.



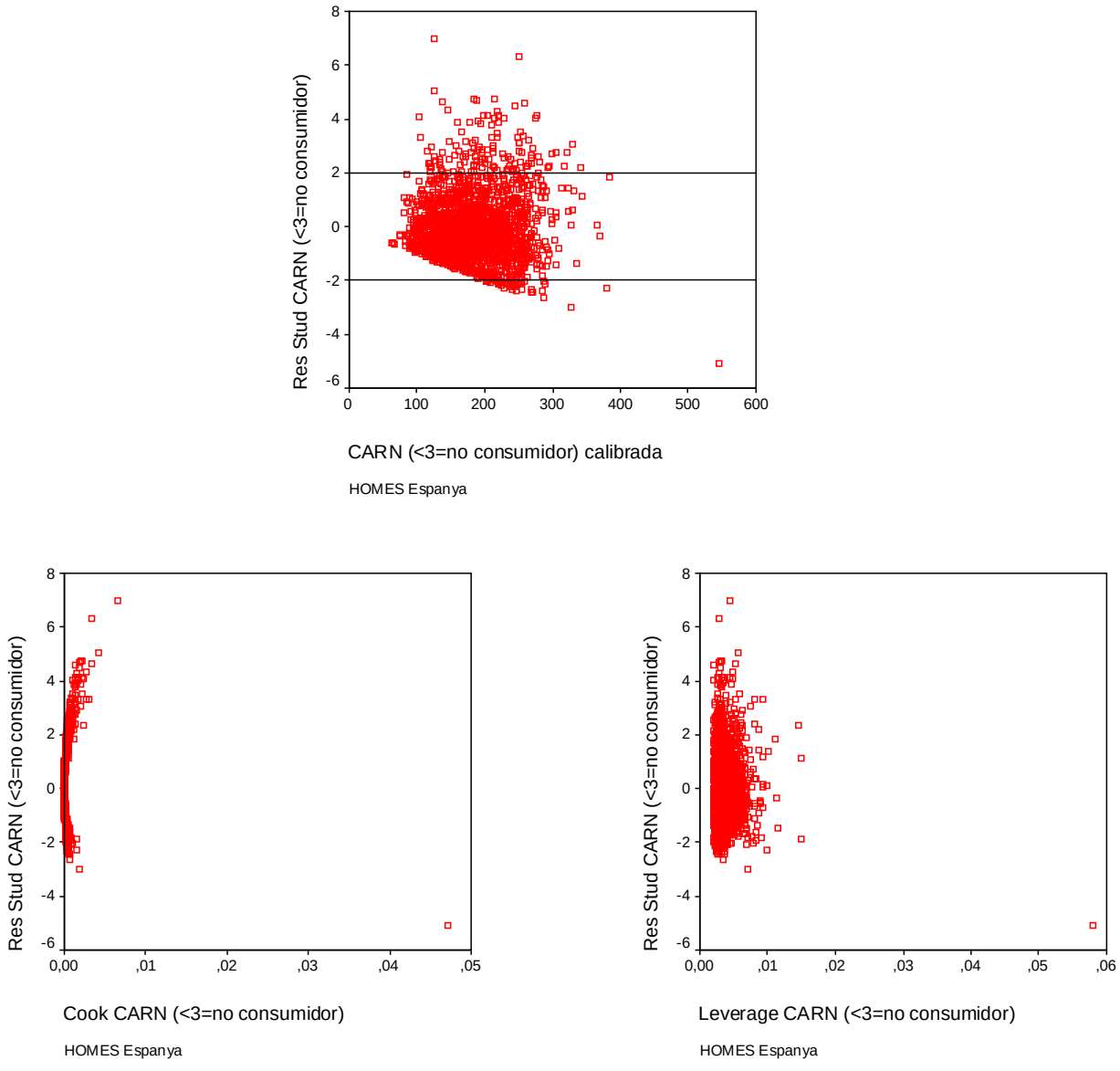






Els punts influents, avaluats amb els coeficients de Cook i el *leverage* també tenen un comportament força similar a l'obtingut amb les dades originals, per a tots els països. Es mostren les dades per als homes espanyols (figura A16).

Figura A16. Diagrames de dispersió dels residus estudentitzats del calibratge de carn respecte a la variable predita (calibrada), coeficient de Cook i *leverage* després d'excloure els consumidors de menys de 3 g/dia en el QFA/HD per als homes d'Espanya.

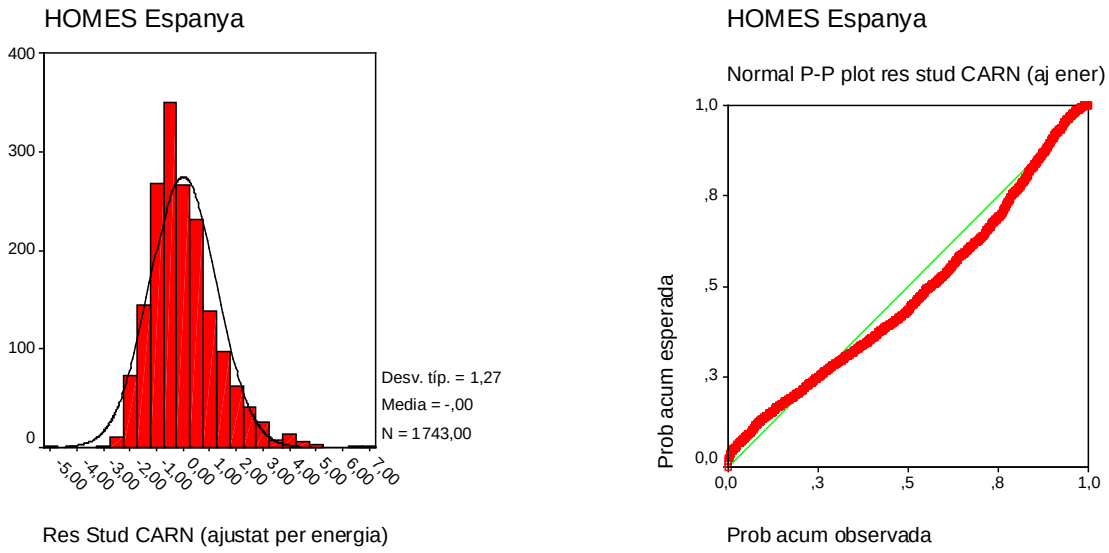


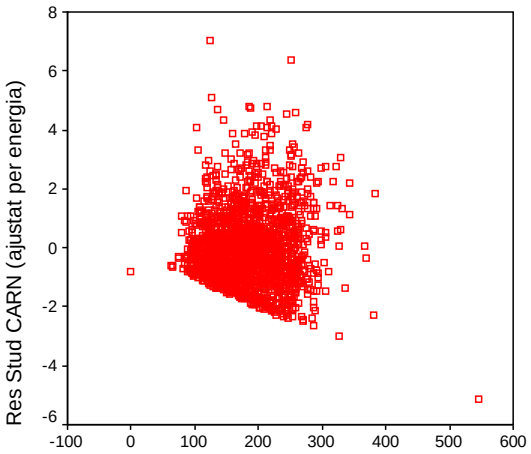
A2.4 AJUST PER ENERGIA

Alguns autors indiquen la necessitat d'ajustar pel consum total calòric (energia) a l'hora de calibrar. Incorporant aquesta variable en el model de calibratge no es modifiquen pràcticament els resultats, tant en homes com en dones. L'r^2 del model no varia més d'una centèsima.

L'ajust del model és pràcticament idèntic ajustant o no per energia, tant en termes de residus, punts influents o *outliers*. Es mostren els gràfics per als homes d'Espanya (figura A17).

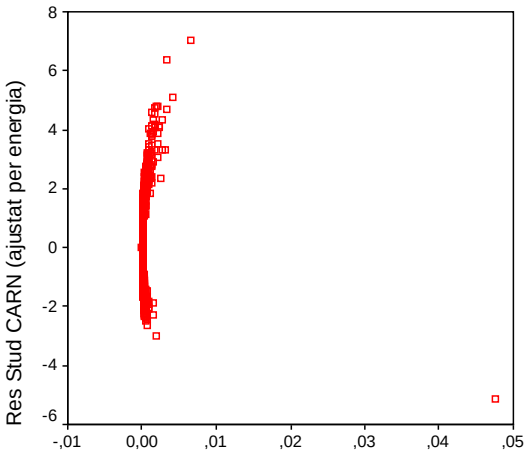
Figura A17. Histograma, dibuix de probabilitat normal i diagrames de dispersió dels residus estudentitzats del calibratge de carn respecte a la variable predita (calibrada), coeficient de Cook, *leverage* i consum de carn QFA/HD en usar un model de calibratge amb l'energia com a covariable per als homes d'Espanya.





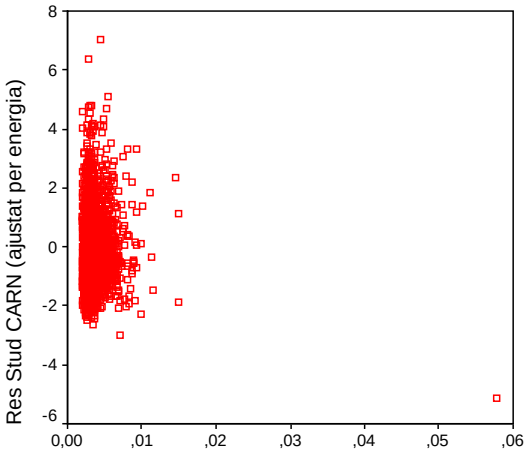
CARN (ajustat per energia) calibrada

HOMES Espanya



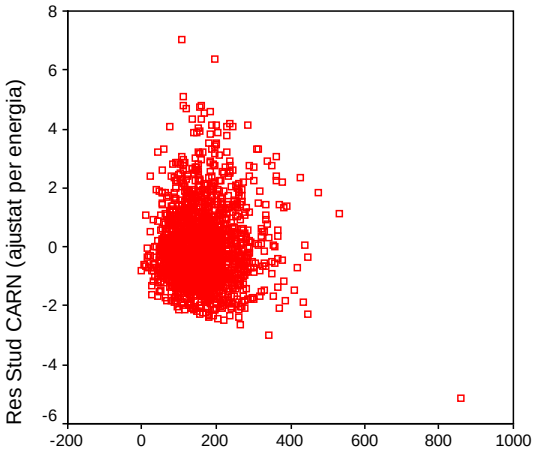
Cook CARN (ajustat per energia)

HOMES Espanya



Leverage CARN (ajustat per energia)

HOMES Espanya



CARN (Q)

HOMES Espanya

A3. ESBORRANY PER A UN ARTICLE

A3.1 INTRODUCCIÓ

L'error de mesura és un dels problemes que sempre ha acompanyat la Ciència. Gairebé cap estudi se'n pot lliurar i, com a màxim, es pot intentar minimitzar-lo perquè la influència sobre les conclusions derivades de l'experiment sigui negligible. L'error de mesura i el control d'aquest també és un dels reptes principals als quals s'enfronta l'Estadística. La qualitat de les mesures és fonamental en qualsevol àmbit, però pren un especial interès en el camp de les Ciències de la Salut, on contínuament es prenen decisions basades en els mesuraments efectuats (White 2003).

En aquest treball em centraré en l'error de mesura de la dieta, obtinguda amb diversos instruments i en diferents poblacions, i en com es pot intentar corregir aquest error quan relacionem la dieta habitual d'un individu amb l'aparició d'una malaltia rara i crònica com és el càncer.

En un estudi etiològic univariant en què volem saber l'efecte d'una certa variable sobre una determinada malaltia, si la variable és mesurada amb error, l'efecte observat és menor que l'efecte real (atenuació de l'efecte) (Greenland 1980, Kupper 1984).

A més, en grans estudis multicèntrics, en què els qüestionaris de dieta acostumen a ser diferents entre centres per capturar les dietes locals (Friedenreich 1994), la magnitud i la natura dels errors sistemàtics i aleatoris pot variar entre els centres i distorsionar l'estimació i interpretació de la relació global entre la dieta i la malaltia quan es combinen les diferents cohorts (Slimani 2002).

Una de les eines més utilitzades per corregir l'error de mesura de la dieta quan calculem un paràmetre que relacioni dieta i malaltia és el calibratge (Stürmer 2002).

El primer objectiu del calibratge és, a nivell individual, intentar corregir el biaix d'atenuació en el risc relatiu (o altres mesures d'associació) degut als errors aleatoris de la mesura de la dieta (Slimani 2002). En estudis multicèntrics el segon objectiu seria, a

nivell ecològic, ajustar per una sobre o infraestimació sistemàtica de la dieta a cada centre.

Aquest treball pretén descriure, justificar i aplicar mètodes de calibratge usant la regressió lineal, per corregir els errors de mesura i estandarditzar les dades provinents de diversos centres relatives al consum de carn en una cohort europea de més de mig milió de persones de 10 països i la relació amb l'aparició de càncer gàstric (CG).

A3.2 POBLACIÓ I MATERIALS

L'estudi EPIC

L'estudi EPIC (*European Prospective Investigation into Cancer and nutrition*) és un estudi prospectiu multicèntric sobre dieta i càncer que inclou 28 cohorts de 10 països d'Europa Occidental (França, Itàlia, Espanya, Regne Unit, Alemanya, Holanda, Grècia, Suècia, Dinamarca i Noruega). Formen la cohort 521.468 persones, destacant l'heterogeneïtat dels patrons dietètics, factors socioculturals i hàbits de vida dels seus participants. El reclutament s'efectuà entre 1991 i 2000. El seguiment mitjà és de 5,0 anys. La informació sobre la dieta habitual (referida a l'any anterior) es va recollir a través d'un qüestionari de freqüència alimentari (QFA) o d'una història de dieta (HD) desenvolupades i validades a cada país. A més, es va administrar un record de 24 hores (R24H) a 36.994 persones per ser usat com a mètode de referència del calibratge.

L'estudi de calibratge

La mostra de calibratge es va seleccionar de forma aleatòria de la mostra principal, ponderada per gènere i edat tenint en compte la distribució per dies de la setmana i estacions de l'any (Slimani 2002). L'estudi de calibratge es va dur a terme entre 1995 i 2000. La taxa de participació va ser alta, amb 7 països per sobre del 75%. Per mirar la representativitat de la mostra de calibratge es van comparar les variables més rellevants respecte a la mostra principal. No es van trobar diferències en general (Slimani 2002).

Mètodes de mesura de la dieta en l'EPIC

El mètode que s'usa com a qüestionari dietètic de referència és el R24H. L'enquestat ha de reportar tot allò que ha menjat durant el dia anterior a l'entrevista, estructurat en 11 ocasions d'ingesta i ajudat per un programa informàtic. El R24H és un mètode obert que permet una descripció força detallada d'un gran nombre de plats i receptes heterogenis (Witschi 1990).

La majoria dels centres de l'EPIC van utilitzar diferents qüestionaris per mesurar la dieta habitual a l'inici de l'estudi. Bàsicament, però, els podem agrupar en QFA i HD. Ambdós mètodes avaluen el consum habitual d'una persona durant l'any anterior. El QFA és una graella on a les files hi ha una llista d'aliments o receptes estàndard. A les columnes, un indicador de freqüència. L'individu ha d'anar omplint la graella amb els aliments que consumeix indicant-ne la freqüència. El mètode HD s'acostuma a contestar amb l'ajuda d'un entrevistador i un programa informàtic. Es repassen totes les ocasions d'ingesta al llarg del dia i es pregunta pel consum habitual al llarg de l'any (en una setmana tipus).

A3.3 MÈTODES ESTADÍSTICS

Fonaments teòrics del calibratge

Bàsicament el calibratge consisteix a aprofitar les dades d'un qüestionari no esbiaixat aplicat a una part de la mostra, per corregir els riscos relatius (RR) (o altres mesures d'associació) basats en les dades aportades per un (o diversos) qüestionaris de menys qualitat, afectes d'error, però aplicats a tota la mostra. El qüestionari general és més fàcil d'administrar i barat que el qüestionari de referència.

Suposem que la taxa d'incidència d'una malaltia (*a*) es relaciona amb el consum habitual d'un aliment, *T,* mitjançant un model log-lineal:

$$\log(a) = \alpha^* + \beta^* T \qquad \text{(I).}$$

Si no disposem de la dieta real *T*, haurem de fer servir un qüestionari (Q), afecte d'error de mesura:

$$Q=\phi_Q+\delta_Q T+\varepsilon_Q \quad \text{(II)}$$

$$\log(a)=\alpha+\beta Q \quad \text{(III)}$$

amb error aleatori ε_Q distribuït normalment amb mitjana zero i variància $\sigma^2_{\varepsilon Q}$ independent de *T*. Els coeficients ϕ_Q i δ_Q representarien el biaix constant (quan un individu tendeix a infra o supraestimar el seu consum de forma constant) i d'escalat proporcional (quan aquesta infra o supraestimació és proporcional al consum real) respectivament.

L'efecte $\hat{\beta}$ estimat en relacionar la taxa de malaltia amb l'exposició mesurada amb el mètode general *Q* ve esbiaixat per un factor λ (anomenat factor de calibratge), així que la relació real entre la malaltia i l'exposició serà $\hat{\beta}^* = \hat{\beta}\hat{\lambda}^{-1}$, havent obtingut $\hat{\lambda}$ de regressar *T* (o un mètode de referència vàlid *R*) respecte de *Q*.

Un procediment alternatiu a haver de calcular $\hat{\beta}$, $\hat{\lambda}$ i haver de dividir-les és calcular directament l'efecte corregit $\hat{\beta}^*$ usant en el model (I) la ingesta calibrada (X) en comptes de *T* (Plummer 1994). Aquesta ingesta calibrada *X* no és res més que els valors predits per a tota la cohort usant la ingesta general *Q* i aplicant els coeficients obtinguts de la regressió entre *T* (o *R*) i *Q*. Es pot demostrar que $\lambda = \rho^2_{QT}/\delta_Q$ (Kaaks 1995a). En estudis dietètics és habitual que $\lambda<1$, per tant també se l'anomena factor de desatenuació, ja que corregeix l'atenuació produïda en β per l'error de mesura.

En estudis dietètics, habitualment no es pot mesurar directament *T*. Per estimar *X* fa falta un estudi addicional amb una mesura de referència *R* sense biaix:

$$R = T + \varepsilon_R \quad \text{(IV)}$$

amb ε_R amb mitjana zero i independent dels errors de mesura del qüestionari general [$\text{Cov}(\varepsilon_R, \varepsilon_Q)=0$]. Si aquests errors són independents, no s'espera que ε_R causi biaix en

l'estimació de λ (aquest vindria donat només per ε_Q); per tant les mesures de referència R no han de ser necessàriament precises i es poden basar en una única administració de la mesura de referència (Kaaks 1997), ja que R no té error sistemàtic. Sota aquestes assumpcions E[R|Q]=E[T|Q] i els valors predits (calibrats) X es poden estimar a partir de la regressió de R sobre Q.

En els estudis multicèntrics, a més, l'heterogeneïtat de l'estimador de β_i per cada cohort deguda a les diferències del biaix provocades pel diferent grau d'error en els qüestionaris Q_i així com la millora de la precisió de l'estimador global β pot ser corregida per l'ús de la mesura de referència R (via calibratge) (Kaaks 1994a).

El calibratge permet corregir simultàniament les diferències entre cohorts quant a biaix d'atenuació (degut als errors aleatoris en els qüestionaris basals) i els biaixos proporcionals d'escala δ_Q deguts a les correlacions intracohort entre els errors de mesura i els valors reals d'ingesta (Kaaks 1995b).

La variància del nou estimador corregit ve donada per (Rosner 1989):

$$Var(\hat{\beta}_i^*) = \frac{1}{\hat{\lambda}_i^2} Var(\hat{\beta}_i) + \frac{\hat{\beta}_i^{*2}}{\hat{\lambda}_i^4} Var(\hat{\lambda}_i) \qquad \text{(V).}$$

Si la mostra de calibratge és prou gran podem assumir que el segon sumand de (V) és gairebé 0.

Com que el biaix d'atenuació depèn només de l'error aleatori de la variable predictora Q no s'espera que els errors aleatoris a la mesura de referència R causin cap biaix en l'estimació de λ. Això justifica que es pugui prendre com a mètode de referència una sola mesura de la dieta mitjançant un mètode no esbiaixat com el R24H, malgrat que pugui ser una estimació molt poc fiable del consum habitual individual. Perquè el calibratge tingui una precisió suficient cal un nombre prou gran d'observacions, ja sigui incrementant la mostra o el nombre de mesures per individu. Kaaks (1995a, 1995b) demostra que si disposem de N observacions a partir de M mesures en Y individus, la precisió s'optimitza agafant M=1 i Y=N. L'únic inconvenient d'aquesta estratègia és

que no podem estimar separadament ρ^2_{QT}, δ_Q i σ^2_T. Tot i així, quan estimem riscos relatius per diferències de consum no és necessari conèixer aquests estimadors individualment.

El calibratge intenta donar una estimació no esbiaixada del consum mig usant una mesura de referència (Plummer 1994). Per aconseguir aquesta estimació no esbiaixada cal:

1. Evitar el biaix de selecció en seleccionar la mostra de calibratge.
2. Què el mètode de referència sigui no esbiaixat.

Com es veu, no cal que el mètode de referència sigui molt fiable ja que l'objectiu és caracteritzar bé la mitjana d'ingesta de la subcohort, no la d'un individu dins una subcohort. Però si el mètode és poc fiable l'error aleatori serà important. Això es pot compensar usant una mostra prou gran (Plummer 1994).

Aplicació del model de calibratge a l'estudi EPIC

El model de calibratge que aplicarem en aquest projecte serà un model de regressió lineal amb efectes fixos. La variable dependent serà el consum de carn mesurat en el R24H. Com a variable explicativa hi haurà la interacció del país amb el consum de carn mesurat amb QFA/HD, de forma que obtindrem coeficients de calibratge diferents per a cada país, i les variables d'ajust: centre, edat al reclutament , alçada, pes i estació de l'any en que es mesura QFA/HD. Els models es correran separats per sexe. El fet d'obtenir coeficients de calibratge per a cada país permet tenir en compte l'especificitat geogràfica en termes de qualitat de mesura de la dieta i rang d'ingesta. No es calculen a nivell de centre en pro d'obtenir coeficients més estables. Per assegurar que tots els dies de l'any hi són igualment representats, es ponderarà per la combinació estació astronòmica–dia de la setmana.

En els models de calibratge s'exclouran aquells que en el QFA/HD han reportat un consum habitual de carn de zero grams (no consumidors), ja que s'assumeix que l'error de mesura en els no consumidors és pràcticament inexistent. A aquests individus se'ls imputarà directament un valor de zero en la variable predita (calibrada).

A part del model original que acabem de presentar, es provaran uns altres models, amb transformacions de les variables d'ingesta (arrel quadrada i logaritme), exclusió dels consumidors per sobre del percentil 99, reagrupament dels consumidors de menys de 3 grams de carn amb els no consumidors i ajust addicional per consum calòric.

Per a cada model es calcularan les prediccions (variable calibrada), i per avaluar l'ajust, els residus estudentitzats i els coeficients de Cook i *leverage*.

Model de malaltia

Per estudiar la relació entre el consum de carn i l'aparició de CG en la nostra cohort usarem un model de riscos proporcionals de Cox (1972). L'edat serà l'eix de temps en els nostres models. Ajustarem per la variable centre a més d'estratificar l'anàlisi per país. Es podria pensar que els individus que van ser diagnosticats en els primers anys de seguiment podrien haver canviat d'hàbits a causa de malalties precursores o símptomes del seu CG. Realitzarem un anàlisi de sensibilitat, estudiant inicialment tots els casos i restringint després l'anàlisi a aquells que han estat seguits almenys 2 anys. Les variables d'ajust que s'utilitzen seran el nivell educatiu, el consum de tabac, l'índex de massa corporal (IMC) i l'energia consumida. Per poc que es pugui es mantindran els homes i dones en el mateix model, pel nombre limitat de casos disponible. Com que al moment de calibrar s'han exclòs els no consumidors i se'ls ha assignat directament un valor zero en la variable calibrada, controlarem aquest fet per mitjà d'una variable indicadora.

En fer el model de Cox, no tenim en compte que la variable calibrada prové d'un model de regressió. Així, el HR estimat té una variància infraestimada. Per corregir aquesta variància es fa un procediment *bootstrap*. *Z* cops es remostreja la mostra que disposa de dades del qüestionari basal i del de referència, amb repetició. Això fa que obtinguem *Z* variables calibrades diferents. Estimem un model de Cox *Z* cops, cadascun amb una de les diferents variables calibrades obtingudes, i per a tota la cohort. Això dóna lloc a una col·lecció de *Z* HR's. Aleshores, podem estimar l'error estàndard (SE) corregit com (Rosner 2001):

$$SE_{corregit}(\hat{\beta}^*) = \sqrt{(\sum_{b=1}^{Z} \text{var}(\hat{\beta}_b^*)/Z) + \frac{1}{Z-1}\sum_{b=1}^{Z}(\hat{\beta}_b^* - \bar{\hat{\beta}}^*)^2} \qquad \text{(VI)}.$$

A3.4 RESULTATS

Descripció de la mostra

Dels 521.468 individus que formen la cohort EPIC 56.059 n'han estat exclosos perquè tenen un CG prevalent, problemes amb les dates de seguiment, perquè no disposen d'un qüestionari de dieta basal, són noruecs (cohort amb seguiment molt curt) o pel fet de tenir una raó energia consumida/energia requerida (EI/ER) per sobre del percentil 99 o per sota del percentil 1; en resta una mostra efectiva per a l'anàlisi de 465.409 individus pertanyents a 9 països (taula 4).

A l'estudi de calibratge hi participaren 36.994 individus, dels quals 2.538 foren exclosos per les mateixes causes que a l'estudi general, i se'n va obtenir una mostra final de 34.456 individus (taula 4). S'han detectat 270 casos de CG.

A la taula 6 es poden veure algunes característiques dels casos i individus censurats. Degut a la presència de valors mancants, per als models de malaltia disposarem de 460.693 individus a risc i 268 casos.

L'estudi de calibratge

Com a mesura de precaució, no es calibraran aquelles variables en els centres en què les mitjanes del qüestionari general i del R24H tenen una raó inferior a 0,5 o superior a 2,0. Això indicaria que probablement els qüestionaris no mesuren el mateix (per exemple, degut a la inclusió d'un aliment força consumit en només un del qüestionaris). Podem veure quines són les mitjanes de consum de carn dels qüestionaris basals i de R24H per centre i gènere, així com la raó de mitjanes (taula 7). Aquesta raó varia entre el 0,62 dels homes d'Umea i el 1,68 de les vegetarianes d'Oxford. Així, el consum de carn té mitjanes prou similars entre els diferents qüestionaris, cosa que permet incloure tots els centres en el model de calibratge.

Si tenim en compte que les mitjanes del R24H ens donen la millor estimació de la ingesta d'un aliment a nivell grupal, a la taula 7 podem veure que el rang de consum de carn entre els centres és molt ampli. Sant Sebastià té el consum més elevat (242 i 127

grams/dia respectivament per homes i dones) mentre que Grècia té el consum més baix (77 i 46 grams/dia respectivament), si no considerem la cohort vegetariana d'Oxford.

Un cop efectuat el calibratge podem comparar els valors de la variable original (QFA/HD), de la variable de referència (R24H) i la variable predita que s'usarà en el model de malaltia (variable calibrada) (taula 8). Podem veure que un dels efectes del calibratge és un "encongiment" de les dades. Com calia esperar, la variable del R24H és la que té més variància, pel fet que es basa en el consum d'un dia. Els extrems de la variable calibrada també són més suaus que els de les variables original i de referència. Cal destacar la presència de valors negatius en la variable calibrada, degut a l'efecte de les covariables. Recordem que als no consumidors (aquells que reporten 0 grams en l'estudi basal se'ls hi assigna un zero directament, però no així als molt baixos consumidors. Aleshores l'efecte, fins i tot lleu, d'una covariable pot portar a prediccions negatives). Cal dir que en aquest cas els valors negatius només es donen a la cohort vegetariana d'Oxford, on els consums són molt baixos.

A la taula 9 es poden veure els coeficients de desatenuació per a cada país amb els intervals de confiança. Cal recordar que cada centre, a més, té un terme independent propi. Per valors de $\lambda<0,2$ o $\lambda>1,0$ els models de calibratge poden ser massa inestables (relació massa dèbil entre les dues variables de mesura de la dieta). Tots els factors de calibratge per carn estan entre 0,27 i 0,73. Són una mica més grans pels homes (mitjana no ponderada de tots els factors de calibratge=0,48) que per les dones (mitjana no ponderada=0,38, 0,39 excloent-ne França). Holanda i Alemanya són els països amb factors de calibratge més alts, tant en homes com en dones (mitjana no ponderada de 0,59) mentre que Dinamarca en homes i Grècia en dones tenen els coeficients menors (0,27). El Regne Unit (mostra de calibratge més baixa) i Grècia tenen els errors estàndard majors.

Ajust del model de calibratge

Pel consum habitual de carn *Q* existeixen cues prou llargues per la dreta. Despreciant aquestes cues, la distribució seria prou normal, excepte al Regne Unit (figura 8). Pel consum de carn mesurat amb el R24H la distribució és clarament asimètrica, ja que acostuma a donar molt de pes als zeros (figura 9). Els gràfics de dispersió entre el consum de carn basal *Q* i el de referència *R* ens donen una idea de la relació entre les

dues mesures (figura 10). L'ajust d'una recta dóna coeficients de r^2 molt baixos (sempre per sota de 0,20, variant entre 0,02 per a les dones de Grècia i 0,18 per als homes britànics). És difícil assumir normalitat dels residus (figura 11). L'efecte de la falta de normalitat és la pèrdua d'eficiència dels estimadors obtinguts. Per tant els tests sobre la significació del paràmetre poden ser no vàlids, però sí que podem limitar-nos a fer una estimació puntual del paràmetre (Peña 2000). Més endavant podem aproximar la l'error estàndard del paràmetre mitjançant *bootstrap*. Es veuen uns residus sense forma aparent, però esbiaixats cap a valors positius (hi ha molts més residus per sobre de 2 que per sota de -2), degut probablement a l'asimetria dels valors extrems (poden haver-hi valors extrems només positius). Hi ha una sèrie reduïda de punts que podrien ser influents com es desprèn de les gràfiques de Cook i *leverage* (figura 12).

Com s'ha pogut observar, l'ajust dels models de calibratge no és gaire bo. Recordem, però, que l'important es disposar d'una estimació puntual; no cal que sigui precisa, però sí que funcioni bé a nivell grupal. Observem doncs què passa si mirem la relació entre la variable original *Q* i la de referència *R* a nivell de centre per a homes i per a dones separadament (figura 13). L'ajust millora de forma espectacular.

Modificacions al model de calibratge original

L'arrel quadrada de la variable QFA/HD és força compatible amb una distribució normal (figura A1). Creuant la variable del QFA/HD amb la del R24H no s'aprecia cap millora important en els coeficients r^2, tant per la transformació arrel quadrada com logaritme. No hi ha millores importants respecte a l'ús de les variables no transformades per a cap país. Si mirem els residus observem una lleugera millora en la normalitat a l'aplicar l'arrel quadrada. Un cop explorades les distribucions de la variable carn (tant del QFA/HD com del R24H), tant usant la variable original, com la transformació arrel quadrada o logaritme, i també el comportament de la variable predita i els seus residus, *outliers* i punts influents, no trobem cap millora significativa pel fet d'aplicar una transformació.

En una mostra gran, com la que tenim en l'estudi, és poc probable que uns pocs valors extrems modifiquin la relació entre variables. Repetirem l'anàlisi per als homes i dones, excloent-ne els qui consumeixen més del percentil 99 aproximadament. La variació en els coeficients de desatenuació respecte a l'ús de les variables sense exclusions és baixa.

La distribució del consum de carn corresponent al QFA/HD és més simètrica i normal. Pels R24H el pes dels no consumidors continua essent molt important. Creuant les dues mesures excloent els grans consumidors no hi ha cap canvi evident en el núvol de punts ni en l'ajust de rectes de regressió. Els residus tenen un comportament gairebé idèntic al que tenien abans d'excloure els grans consumidors.

Poden aparèixer valors molt petits de consum provinents de receptes (habitualment receptes estàndard, que l'enquestat no pot modificar). Podem intentar reclassificar els consumidors de quantitats petites com a no consumidors. Hem triat consumidors habituals de menys de 3 grams al dia de carn com a "no consumidors". Els coeficients de desatenuació gairebé no varien al excloure els consumidors de menys de 3 g/dia i l'ajust del model de calibratge no varia gaire al excloure als baixos consumidors.

Alguns autors indiquen la necessitat d'ajustar pel consum total calòric (energia) a l'hora de calibrar. Afegint aquesta variable en el model de calibratge no es modifiquen pràcticament els resultats.

Aplicació de les dades calibrades a un model de Cox per CG

La variable carn calibrada és expressada de forma contínua, en consum de 100 grams/dia. Es proven 2 models separats per gènere, estratificats per país i ajustats per centre, consum de tabac, IMC, nivell educatiu i ingesta energètica.

La diferència en els HR de carn entre homes i dones és de més d'un punt (1,80 els homes i 3,24 les dones, només significativament diferent d'1 per aquestes últimes) (taula 10). Provant un model amb la variable calibrada, les 4 variables d'ajust i la variable gènere, així com la interacció d'aquesta amb la variable calibrada i comparant-lo amb el mateix model sense el terme d'interacció obtenim una diferència en la log-versemblança d'1,38 amb un grau de llibertat, cosa que indica que el model amb interacció no és significativament millor que el que no té aquest terme. Així doncs, el model que s'utilitzarà serà el conjunt per homes i dones amb una variable indicadora que els diferenciï.

Els HR obtinguts amb el model conjunt (taula 11) estan entre mig dels obtinguts amb els models separats per gènere, però amb més significació ja que es guanya potència en tenir més casos. El tabac continua essent un factor de risc (HR=1,34 i 1,72 per ex-fumadors i fumadors actuals respectivament, p conjunta=0,003), mentre que no s'observa cap efecte global del nivell d'estudis (p conjunta=0,44). L'IMC apareix protector (HR=0,96, p=0,026) i el consum calòric no mostra cap efecte (HR=1,00, p=0,21). Les dones tenen menys risc de patir CG (HR=0,68, p=0,032). Per últim, la nostra variable d'interès, la carn calibrada, és un factor de risc de CG (HR=1,97, IC95%=1,21-3,22).

La millora aportada pel calibratge la podem estimar comparant els resultats obtinguts amb les variables calibrada i sense calibrar. Els resultats abans de calibrar no són gaire diferents (taula 12). Com hom podia esperar el valor del HR per a la variable sense calibrar és una mica menor (HR=1,43) i té un interval de confiança més estret (IC95%=1,13-1,81).

Ajust del model de Cox

El model assumeix riscos proporcionals. Els residus de Schoenfeld, específics per a cada variable predictora, creuats amb el temps de seguiment poden ajudar a detectar variables que no compleixin aquest supòsit. Esperem no trobar cap tendència (Therneau 2001). No es detecta cap tendència identificable (figura 14). Els residus de martingala permeten veure si el model ajusta bé (Therneau 2001). En la figura 15, en què es creua el predictor lineal (resultant d'emprar totes les covariables del model) amb els residus de martingala veiem que el model no classifica bé els casos, ja que s'espera que els residus de martingala estiguin al voltant de zero. Això és degut a què el nombre de casos (en color verd) és ínfim comparat al d'individus a risc (en vermell) i per tant l'error global de classificació és molt baix tot i classificar malament a tots els casos. Repetint el mateix gràfic per a cadascuna de les variables predictores individualment el resultat és anàleg.

Recordem, però, que estem més interessats en buscar una associació entre CG i carn que en fer prediccions. Per tant, una hipòtesi a comprovar és la linealitat de l'efecte. Repetint el mateix gràfic anterior dels residus de martingala, però ara respecte al consum de carn, si fem un suavitzat del diagrama de dispersió esperarem trobar una

corba aproximadament plana respecte al consum de carn. Si no trobem aquesta corba voldrà dir que la forma funcional de la variable independent (carn) no és acceptable. Com es veu a la figura 16 obtenim una línia gairebé recta a l'alçada dels individus a risc, que ens permet acceptar l'assumpció de linealitat de l'efecte de la carn calibrada. Una anàlisi dels punts influents, mitjançant l'estadístic LD (desplaçament de la versemblança en eliminar un punt) (SAS 2001) mostra cinc casos com a possibles punts influents (figura 17). Eliminant aquestes cinc observacions es modifiquen lleugerament els coeficients obtinguts. Podem concloure que l'ajust del model utilitzat és raonablement acceptable.

En estratificar per país assumim que entre aquests hi ha un cert efecte homogeni del consum de carn sobre el CG. Si bé observem a la taula 13 diferències entre països (el HR per 100 grams de carn varia de 0,83 a Espanya a 2,17 a França si usem la variable original i de 0,42 a Grècia a 13,60 a França si usem la variable calibrada), els estimadors són molt imprecisos, cosa que fa que en fer un test d'heterogeneïtat no obtinguem resultats significatius (obtenim una khi-quadrat amb 8 graus de llibertat de 7,25 per a les dades originals i de 6,22 per a les dades calibrades, en comparar la versemblança entre models amb i sense termes d'interacció país-consum de carn).

Un dels possibles problemes quan avaluem la relació entre el consum de carn i el CG és que els individus als quals se'ls ha diagnosticat un càncer en els primers mesos de seguiment podrien haver canviat els seus hàbits dietètics (i altres, com el consum de tabac) precisament perquè ja tenien la malaltia (o una precursora d'aquesta) però encara no havien estat diagnosticats. Repetint l'anàlisi excloent els casos i individus censurats que tenen un seguiment menor de 2 anys el HR per carn (x100 g.) creix fins 2,30, desapareix l'efecte de l'IMC i augmenta el del tabac (taula 14).

La variància estimada pel model de Cox està infraestimada, ja que no té en compte la variabilitat que el model de calibratge aporta. Per resoldre aquesta situació es fa un procediment *bootstrap*, que calcula 300 vegades el HR del model de Cox, a partir de 300 estimacions de la variable carn calibrada. L'error estàndard del log(HR) sense fer *bootstrap* era de 0,0025 grams/dia. L'error estàndard corregit val 0,002621 grams/dia, o sigui un 4,84% més. D'aquesta forma els intervals de confiança originals del HR (1,21-3,22) es corregirien a (1,18-3,30).

Usant uns altres models en què s'usa la variable transformada o amb restriccions, tant per calibrar com pel model de risc s'observa sempre un efecte de risc per la carn (taula 15).

A3.5 DISCUSSIÓ

Si comparem els nostres resultats amb els d'altres estudis, els valors d'r^2 del model de calibratge s'assemblen als obtinguts en uns altres estudis usant el R24H com a eina de referència. Per exemple, Rosner (2001) obtingué un r^2 per carn de 0,06 mentre que aquí observem valors entre 0,02 i 0,18 (segons país). No és clar que la carn en general sigui un factor de risc per CG si ens atenem a la literatura (WCRF 1997). En qualsevol cas les mateixes fonts sí que citen alguns tipus de coccions i preparacions (embotits, carn fumada o curada) de la carn com a factors de risc de CG. Ward (1997) troba un OR de 2,4 per als qui consumien més de 19 cops a la setmana carn vermella envers els qui en consumien menys de 8 cops. Així, és probable que el HR d'1,97 que hem trobat pugui ser més gran si ens centrem en alguns tipus concrets de carn o en com es prepara, conserva o cuina.

A continuació es discuteixen cadascuna de les assumpcions en què es basa la teoria del calibratge i l'aplicació de la variable calibrada obtinguda en un model de Cox:

***R* no esbiaixada:** la mesura de la dieta *R* en el subestudi de calibratge només cal que sigui no esbiaixada (respecte a la realitat *T*), més que un reflex de la ingesta absoluta. Per tant, es poden obtenir estimadors no esbiaixats de λ usant una mesura única de la dieta (com el R24H) (Rosner 1990). Aquesta reducció del nombre de mesures, però, comportarà un increment dels errors estàndard de λ i per tant uns intervals de confiança més amples per β*. Una solució seria incrementar el nombre de subjectes participants a l'estudi de calibratge (Rosner 1988). La participació en l'estudi de calibratge ha de ser alta per evitar biaixos de selecció indesitjables. En l'estudi EPIC, com hem vist, la participació va ser prou alta (superior al 75% en 7 països). En aquest projecte no és pot demostrar si la mesura de referència és o no esbiaixada en no disposar d'una tercera mesura de referència no correlacionada amb els errors d'*R* i *Q* i no esbiaixada (marcador

bioquímic). Però, fins i tot si *R* fos esbiaixada però la direcció i magnitud de l'error fos aproximadament igual per les diferents subcohorts, la mesura de referència encara es podria usar per fer un calibratge a nivell ecològic (calibratge entre cohorts o estandarditzar les mesures per obtenir un estimador comú). Això vol dir que les mesures calibrades no tindrien validesa absoluta, però si relativa. Fins i tot seria vàlida per corregir a nivell individual si l'error fos homogeni dins de cada grup (Kaaks 1997, Riboli 2000).

Independència entre els errors de *Q* i *R* (donat *T*): és difícil d'assumir en general (Kaaks 1995a) i impossible de provar en aquest projecte. Qüestionaris en què s'usa la memòria de l'individu per contestar poden tenir errors correlacionats, sobretot si s'administren en períodes de temps molt pròxims (Freedman 1991). Si això passa, la variància del consum estimat *X* pot estar sobreestimada. En l'EPIC els dos qüestionaris utilitzats es basen en respostes a preguntes, per tant s'utilitza la memòria, i existeix heterogeneïtat quant al temps transcorregut entre la mesura del R24H i el QFA/HD en la mostra de calibratge. Una possible solució seria l'ús de marcadors bioquímics com a mètode de referència, però lamentablement no es disposa d'un marcador per a cada variable nutricional (Hunter 1990). Uns altres estudis (Cameron 1988) diuen que els mecanismes memorístics per recordar el consum del darrer dia (com el R24H) o a llarg termini (com QFA/HD) són diferents, fet que implicaria una menor correlació entre els errors. Si la correlació entre els errors és negativa l'estimador de RR (o HR) pot estar sobrecorregit (per sobre del valor real) (Wacholder 1993). Tot i així sembla força improbable que els errors de dos mètodes de mesura basats en qüestionaris tinguin correlació negativa. Spiegelman (1997a) demostra com la no correlació entre els errors d'*R* i *Q* dóna lloc a estimadors no esbiaixats de RR (suposant que e_R té mitjana 0 i cov(e_R,T)=0). També mostra com el biaix relatiu de l'estimador corregit del RR varia amb la correlació entre els errors d'*R* i *Q* , amb la fiabilitat d'*R* (defineix fiabilitat=var(R)/var (T)) i amb la correlació entre *R* i *Q*. Concretament l'error d'atenuació del RR (tendència cap a RR=1) s'incrementa a mesura que incrementem la corr(e_R,e_Q), o quan disminueix la fiabilitat d'*R* o quan disminueix la qualitat de *Q* (disminueix corr (Q, T)). Quan la fiabilitat d'*R* és 100% o la corr(e_R,e_Q)=0 o corr(Q,T)=1 no hi ha biaix en l'estimador de RR. Es veu que sempre que corr(e_R,e_Q) no sigui negativa és millor calibrar.

Kipnis (2001) critica el model de calibratge que hem usat (per tenir massa assumpcions, la violació de les quals rebaixaria la potència) però en la proposta que fa calen mesures repetides. Conclou que si s'ignora el biaix individual en l'instrument de mesura, el coeficient de desatenuació estarà més a prop d'1 que allò que realment val (o sigui, no es té en compte tota l'atenuació). Per tant, en qualsevol cas l'estimador obtingut en calibrar d'aquest projecte serà més proper al real que si no calibrem, si bé la correcció efectuada pot ser insuficient per arribar al coeficient real.

Variable contínua: la variable considerada (consum de carn) es pot considerar contínua, ja que no s'aprecien agrupacions de valors que la facin considerar discreta.

Relació *Q*-*T* lineal: no s'observa una relació lineal entre *Q* i *R* (s'usa *R* en no disposar de *T*): els r^2 són sempre menors que 0,20. Però ja sabem que *R* no és un bon estimador de *T* a nivell individual sinó a nivell grupal (país). En aquest cas r^2 puja a 0,61 en homes i 0,57 en dones. Regressions quadràtiques i cúbiques aporten millores poc importants (per exemple, r^2 de 0,65 i 0,66 respectivament en homes).

Normalitat: la normalitat conjunta del consum real i del consum mesurat no sembla ser tan assumible, ja que moltes mesures dietètiques tenen cues que s'aproximen a una distribució log-normal, cosa que provocaria un major error aleatori per consums alts. L'ús de models no paramètrics (o semiparamètrics) pot resoldre el problema (Carroll 1991). Aquest mateix autor arriba a dir que l'assumpció de normalitat no és necessària (Carroll 1990). En aquest projecte la normalitat dels residus observats és discutible, sobretot en alguns centres com el Regne Unit o Grècia. Com veurem, l'ús de transformacions per millorar la normalitat suposa incomplir una altra assumpció: E[R]=T (Boshuizen 2004). L'efecte de la falta de normalitat és la pèrdua d'eficiència dels estimadors obtinguts. Per tant els tests sobre la significació del paràmetre poden ser no vàlids, però sí que podem limitar-nos a fer una estimació puntual del paràmetre (Peña 2000). Més endavant podem aproximar l'error estàndard del paràmetre mitjançant *bootstrap*. La normalitat de la ingesta real o mesurada no sembla tan important, com diu Carroll (1990). En qualsevol cas, les mesures observades en aquest qüestionari no són normals, ja que tenen una important cua per la dreta (Q) i un elevat percentatge de no consumidors (R).

Independència dels errors de mesura d'altres característiques dels individus: els errors de mesura existents han de ser independents entre ells, de la ingesta real i d'altres característiques dels individus (Prentice 1996). Això no és gaire clar per variables com l'IMC. Usant mètodes de laboratori com a referència, s'ha trobat que les persones obeses (quart quartil d'IMC) infrareporten fins un 30-40% de l'energia consumida respecte a les del primer quartil d'IMC (Heittman 1995). Ferrari (2002) mostra com els individus amb major IMC de l'estudi de calibratge EPIC tendeixen a infrareportar el seu consum energètic i els individus amb menor IMC a sobrereportar-lo. El percentatge d'individus identificats com a infrareportadors (aquells que consumeixen menys energia de la que el seu cos requereix segons la fórmula de Goldberg (1991)) a l'EPIC és per sota del 13% en tots els centres (excepte Grècia 20%) en homes i del 17% (Grècia 33%) en dones. El grau d'infrareport és heterogeni entre països però homogeni entre centres d'un mateix país. Cal dir, però, que aquestes estimacions s'han fet assumint que l'activitat física és constant per a tota la població, i potser els obesos són menys actius i necessiten menys energia. S'han proposat alguns models per corregir aquest biaix (Prentice 1996) basats en correccions a nivell individual que necessiten mesures repetides o l'ús de factors de calibratge específics per cada nivell d'IMC. En qualsevol cas no es pot provar aquesta assumpció en aquest projecte en no poder avaluar l'error de mesura de *R*.

Altres assumpcions dels residus del model de calibratge: s'assumeix que no hi ha d'haver una relació dels residus amb *T*. Si mirem la figura 21, en què es creuen els residus estudentitzats amb una estimació de *T*, els valors *R* del qüestionari de referència, no sembla que aquesta assumpció sigui acceptable. Però si tenim en compte que la variabilitat explicada pel model de calibratge és molt baixa, això vol dir que tot allò que no pot explicar el model de calibratge queda dipositat en el residu. D'aquí aquesta forta tendència dels residus envers el consum de referència. Hem d'adonar-nos, però, que aquesta referència no és vàlida a nivell individual (d'aquí el mal ajust), però sí a nivell grupal (país). Per tant, hem de fixar-nos en què passa quan usem dades agregades (mitjanes) per país. Com es veu a la figura 22, tant per a homes com per a dones, no existeix cap tendència dels residus envers el consum del qüestionari del R24H a nivell de país.

Error no diferencial: donat que estem estudiant una cohort, en que tots els individus estan lliures de malaltia a l'inici de l'estudi (que és quan es mesura l'exposició) sembla difícil que pugui existir error diferencial en els qüestionaris *Q* i *R* (o sigui, que el grau d'error depengui del fet de si l'individu acabarà emmalaltint o no) (Rosner 1990). Tot i així, alguns símptomes i malalties precursors de CG poden fer variar la dieta i també la qualitat de les respostes (per exemple, se sap que els individus malalts tendeixen a recordar millor el que han menjat perquè els afecta més). L'exclusió dels individus diagnosticats en els dos primers anys donaria uns resultats en que es pot descartar definitivament la presència d'error diferencial. L'estimador que hem obtingut és major (més llunyà de la hipòtesi nul·la) quan excloem aquests individus diagnosticats prematurament.

Malaltia rara: podem definir com a malaltia rara aquella que té una prevalença baixa i és poc recurrent. En el cas de l'EPIC la prevalença de CG en el moment de fer aquest projecte era de 0,09%. En estudis de simulació, s'ha demostrat que $\hat{\beta}^*$ té poc biaix i una cobertura de probabilitat (percentatge dels cops en què l'interval al 95% de confiança de $\hat{\beta}^*$ inclou el veritable valor de β^*) apropiada si els OR no són majors de 3 en estudis de cohort i la malaltia d'estudi té una prevalença de fins al 5% (Rosner 1989).

Risc moderat: hem observat HR's al voltant de 2, que es poden considerar moderats.

Ingesta constant en el transcurs del temps: aquesta assumpció és impossible de verificar amb aquest estudi. Tot i així, és molt habitual en estudis de dieta en persones adultes.

Error de mesura baix: l'error de mesura del qüestionari de referència no és estimable com ja s'ha dit. El del qüestionari original es situa dins dels límits acceptables amb coeficients de calibratge entre 0,27 i 0,73 segons el país (habitualment s'accepten com a raonables coeficients entre 0,2 i 1).

La relació entre *Q* i *T* ha de ser igual en la mostra de calibratge que en la resta de la cohort: com que la mostra de calibratge s'ha recollit de forma aleatòria, estratificant

per edat i sexe, i tenint en compte l'estació de l'any i dia de la setmana, sembla probable que l'assumpció es compleixi.

$\hat{\beta}$ i $\hat{\lambda}$són independents: aquesta assumpció seria sempre certa si uséssim mostres independents per estimar-les (Rosner 1990) (o sigui, que la subpoblació de calibratge no pertanyés a la població de l'estudi principal). Això no s'ha fet així a l'EPIC però una forma de demostrar que aquesta assumpció és correcta és comparant les β^*'s obtingudes amb tota la cohort amb les obtingudes excloent de l'estudi principal a les persones que van participar a l'estudi de calibratge. S'esperaria que els resultats no variessin pràcticament. En aquest cas el HR varia d'1,97 (p=0,007) a 1,91 (p=0,013) al excloure la mostra de calibratge.

Mètodes de detecció dels individus malalts similars: diferències quant a completitud en la identificació de casos o l'elecció de la data de censura poden crear biaix a nivell intercohorts. En estratificar per país i ajustar per centre ens assegurem que diferències entre aquests no creen biaix a nivell intercohorts (Kaaks 1994a).

Pèrdues pel seguiment independents de l'exposició: aquesta hipòtesi no és demostrable per a aquest projecte. Tot i així és possible que algunes morts tinguin relació amb la dieta (altres càncers, malalties cardiovasculars,...). Repetint l'anàlisi excloent-ne els morts per causes diferents a les de CG obtenim un HR d'1,96 (p=0,007), pràcticament idèntic a l'observat en l'anàlisi amb tota la cohort, cosa que sembla descartar qualsevol efecte en el HR degut a pèrdues pel seguiment relacionades amb l'exposició.

Linealitat: En aquest estudi, hem observat com els residus de martingala són compatibles amb un model d'efecte lineal de la carn.

En aquest estudi s'ha usat el R24H per calibrar. Kipnis (2002) adverteix, però, que l'ús de qüestionaris per calibrar no és el més apropiat. El model que proposa utilitza mesures repetides dels qüestionaris *Q*, *R* i mesures bioquímiques. El model de Kipnis permet correlació entre els errors de mesura del qüestionari. El model que hem usat seria un

model particular d'aquest, assumint *R* no esbiaixada. La presència de biaix relacionat amb el consum real o la correlació entre els errors dels qüestionaris invalidaria el mètode de referència pel model que usem. Com ja hem dit, però, el model de Kipnis no és identificable sense una tercera mesura no esbiaixada i amb errors no correlacionats amb els qüestionaris (biomarcador), i diverses mesures de *Q* i *R*. Cal dir que aquests biomarcadors han de tenir una relació coneguda amb la ingesta real, independent de la quantitat total de menjar ingerida i d'altres trets individuals com l'edat, el gènere, l'IMC,... (Kaaks 1995a). Per ara hi ha molt pocs marcadors que compleixin aquestes condicions. Fins i tot cal que el biomarcador es mesuri en un moment diferent al del qüestionari per evitar la correlació entre els errors aleatoris. Kipnis (2002) mostra com el seu mètode sempre troba coeficients de desatenuació menors que el mètode que hem fet servir (per tant el risc relatiu estimat pel nostre model seria menor que el real, d'un 50% a un 200%), almenys així ho proven amb l'estudi d'ingesta de proteïnes, usant marcadors de nitrogen a l'orina. O sigui, que el mètode que usem sobreestimaria la correlació entre *Q* i *T* (diria que *Q* és més bo del que en realitat és). Les pendents d'*R* amb *T* varien entre 0,34 i 0,77 a l'estudi de Kipnis (quan el nostre model les assumeix 1). La variància del biaix individual d'*R* que assumim 0 varia entre 0,01 i 0,05 a l'estudi de Kipnis, i la correlació entre els biaixos individuals dels qüestionaris *Q* i *R* (que també assumim 0) varia entre 0,35 i 0,95. No és clar si aquests resultats són extrapolables a uns altres nutrients o a proteïnes ajustades per energia (usant també un biomarcador per energia). Per tant, la correcció que efectuem en aquest projecte pot ser incompleta (però en qualsevol cas millor que no fer-hi cap correcció).

Finalment no hem fet cap transformació a les variables de consum de carn. La interpretació dels resultats serà sempre més fàcil si podem usar la variable original. A més cal tenir en compte que la mesura del R24H no és una estimació del consum real a nivell individual sinó a nivell grupal (o sigui E[R]=T, i no R=T). Això té més implicacions de les que pot semblar a primera vista. Per exemple, si regressem *R* en *Q*, tenim E(R|Q). En el model de malaltia acabarem substituint la variable original *Q* per la variable calibrada E[R|Q][= E[T|Q], ja que E[R]=T. Si apliquem l'arrel quadrada i regressem $\sqrt{R}$ en $\sqrt{Q}$ tenim E[$\sqrt{R}$|$\sqrt{Q}$], però, estem, implícitament, assumint que E[$\sqrt{R}$]=$\sqrt{T}$, o el que és el mateix, T= (E[$\sqrt{R}$])^2, que és diferent de l'assumpció bàsica E[R]=T (Boshuizen 2004).

En aquest projecte s'ha usat el model de calibratge de regressió lineal amb efectes fixos. En estudis anteriors, hem provat models multinivell (o sigui, d'efectes mixtos usant el centre com a factor aleatori). Els resultats per als HR obtinguts amb els dos models són força similars. L'ús de mètodes de calibratge més rudimentaris, com el calibratge additiu o multiplicatiu no permet tenir en compte variables de confusió (no permet corregir uns altres errors que no siguin els sistemàtics propis de cada centre o país). Per tant no semblen aconsellables, excepte com a possible solució per al calibratge de grups alimentaris amb un alt percentatge de no consumidors. Per últim, l'ús de mètodes de calibratge no lineals podria millorar la predicció de les variables calibrades, però calen almenys dues mesures del R24H per poder aplicar aquests mètodes (Hoffmann 2002).

Hem vist com l'ajust del model de Cox és prou acceptable. Repetint els anàlisis amb models paramètrics (Weibull, log-normal) els resultats obtinguts són equivalents.

A3.6 CONCLUSIÓ

Al llarg d'aquest document he presentat i aplicat el mètode del calibratge mitjançant una regressió lineal com a mètode d'estandarditzación i correcció dels biaixos dels HR obtinguts amb un model de Cox que relaciona el consum de carn i el CG en un estudi multicèntric.

Com hem vist és suficient emprar una sola mesura de referència en una submostra de la cohort per obtenir prediccions de consum per a tota la cohort, que un cop usades en el model de Cox ens donarà estimadors corregits del HR. Malgrat que el calibratge mitjançant l'ús d'un únic R24H com a mesura de referència pot violar alguna de les assumpcions bàsiques del calibratge (no correlació dels errors dels qüestionaris general i de referència, estimacions no esbiaixades del consum real a partir del qüestionari de referència), que no es poden provar amb les dades disponibles per aquest projecte, el calibratge serveix per corregir almenys una part del biaix amb què estimaríem els HR's si uséssim només les dades del qüestionari general.

És aconsellable l'ús de biomarcadors quan sigui possible, i l'estudi més acurat de mètodes alternatius que permetin emprar variables categòriques, un alt percentatge de no consumidors, relacions no lineals o presència d'errors correlacionats i biaix en els instruments de mesura.

Amb les dades disponibles hem vist com el consum de carn suposa un increment apreciable del risc de patir CG, independentment dels diferents tractaments a què hem sotmès les variables, i de com el calibratge allunyava el HR obtingut de la hipòtesi nul·la.

En definitiva, el calibratge és un recurs per minvar els efectes de l'error de mesura de la dieta en l'estimació dels paràmetres d'associació d'aquesta amb la malaltia.

A4. PROGRAMES INFORMÀTICS

Per realitzar l'anàlisi s'han usat els següents paquets estadístics:

- SAS Release 8.02
- SPSS para Windows Versión 11.0.1
- Stata/SE 8.0 for Windows
- R 1.9.0

Bàsicament el SAS ha estat el paquet més utilitzat, però de vegades s'han aprofitat els procediments dels altres paquets, sobretot per fer gràfics. A continuació es pot veure el codi per obtenir els resultats d'aquest projecte. Per economitzar espai, només es mostren els programes en SAS i Stata, ja que l'SPSS i l'R només s'han fet servir per fer gràfics. Els comentaris estan en color verd.

* PROJECTE.SAS

```
* Projecte de fi de carrera LCTE
* Programa en SAS per fer estudi de calibratge del consum de carn i la relació d'aquest amb càncer gàstric
* Creació 15-03-2004 Modificació 18-10-2004
* per Guillem Pera;

options pagesize=500 linesize=100 nonumber nodate;run;

* La llibreria ORIGIN és d'on surten les dades que s'utilitzen pel projecte EUR-GAST, que es modifiquen i
conserven per aquest projecte a la llibreria DAT;

libname DAT "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades";
libname ORIGIN "C:\EURGAST\Dades\oct02";
libname library "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades";
run;


*************************************************************************************************
*                                PREPARACIÓ DE LES BASES DE DADES                               *
*************************************************************************************************;


* IMPORTACIÓ DE LES DADES:
* PROJECTE: dades baseline  MIDA=521.468 persones
* R24H:     dades del R24H  MIDA=36.994 persones;

data dat.PROJECTE;set origin.ORIGINAL;
         keep IDEPIC IDQST COUNTRY CENTER CNTR_C SEX D_BIRTH D_BIRTH2 D_RECRUI D_DTQST D_NDTQST D_CENSOR
         D_CHECK D_DEATH AGE_RECR  VIT_STAT DATESTOM--CASESTOM L_SCHOOL SCHOOL A_SCHOOL A_SCHO_C WEIGHT_C
         HEIGHT_C BMI CANCER TYP_DIET TYP_DTQ QENER QPROT QCARBO QLIPID QALCOHOL EXCLEIER QG07 SMOKE
         SMOKE_ST A_SMOKE A_SMOK_C A_GIVSM A_GIVS_C FA_INT1--TA_INT5 GIV_SMOK GIV_SMYE TRY_SMOK OCC_SMOK
         N_OCC_SM I_OCC_SM CIGARETT A_CIGRET A_GIVCT N_CIGRET N_CIG_C N_CIG_C2 F_CIGRET--L_CIGRET T_CIGRET
         N_CIG20 N_CIG20C N_CI20C2 N_CIG30 N_CIG30C N_CI30C2 N_CIG40 N_CIG40C N_CI40C2 N_CIG50 N_CIG50C
         N_CI50C2 F_CIG20 F_CIG30 F_CIG40 F_CIG50 CIGARS A_CIGARS A_GIVCS N_CIGARS I_CIGARS T_CIGARS N_CS20
         N_CS30 N_CS40 N_CS50 PIPE A_PIPE A_GIVPI N_PIPE I_PIPE N_PIPE_T N_PI20 N_PI30 N_PI40 N_PI50;
run;

data dat.R24H;set origin.R24_GRP;
         keep IDQST CENTER REC_DAY SEASONS4 RENER RALCO RG07;
run;

data dat.PROJECTE;set dat.PROJECTE;
         length SEX SMOKE--N_CI50C2 SMOKE_ST EXCLEIER BEHASTOM--CASESTOM 3;
run;

* ELIMINACIÓ DELS CASOS PREVALENTS;
data dat.PROJECTE;set dat.PROJECTE;
         if CASESTOM=1 then delete;* s'eliminen 133 persones;
         CASESTO=CASESTOM;
         if CASESTOM=2 then CASESTO=1;
         drop CASESTOM;
         length CASESTO 3;
         label CASESTO='STOMACH CANCER STATUS';
run;

* CREACIÓ DE DATES I IMPUTACIÓ DE DATES MISSING;
data dat.PROJECTE;set dat.PROJECTE;
         * millora de la data de diagnòstic;
         DIAGSTOM=.;
         if (substr(DATESTOM,4,2) in ('00','99')) then DIAGSTOM=input('0107'||substr(DATESTOM,7,4),DDMMYY8.);
         else
                  if (substr(DATESTOM,1,2) in ('00','99')) then DIAGSTOM=input('15'||substr(DATESTOM,4,2)||substr(DATESTOM,7,4),DDMMYY8);
                  else DIAGSTOM=input(substr(DATESTOM,1,2)||substr(DATESTOM,4,2)||substr(DATESTOM,7,4),DDMMYY8.);

         * millora de la data de mort;
         DEATHDAY=.;
         if (substr(D_DEATH,4,2)='00') then DEATHDAY=input('0107'||substr(D_DEATH,7,4),DDMMYY8.);
         else
                  if (substr(D_DEATH,1,2)='00') then
                  DEATHDAY=input('15'||substr(D_DEATH,4,2)||substr(D_DEATH,7,4),DDMMYY8.);
                  else DEATHDAY=input(substr(D_DEATH,1,2)||substr(D_DEATH,4,2)||substr(D_DEATH,7,4),DDMMYY8.);

         * data final del seguiment;
         ENDATE=D_CENSOR;
         if VIT_STAT in (2,6,7,8,9) then ENDATE=min(DEATHDAY,D_CHECK,D_CENSOR);
         if VIT_STAT=1 then ENDATE=min(D_CHECK,D_CENSOR);
         if CASESTO=1 and DIAGSTOM<=ENDATE and DIAGSTOM^=. then ENDATE=DIAGSTOM;
         if CASESTO=1 and DIAGSTOM>ENDATE and DIAGSTOM<D_CENSOR then ENDATE=DIAGSTOM;

         * conversió dels casos post-censoring a controls;
         if CASESTO=1 and DIAGSTOM>ENDATE then CASESTO=0;

         * càlcul de la llargada del follow-up;
```

```
        LENGTH=ENDATE-D_RECRUI;

        * 2305 individus (2301 de Grècia i 4 de Bilthoven) sense dates de censor ni check ni mort (cap d'ells cas);
        * 674 individus amb seguiment=0 i 1 amb seguiment negatiu (cap d'ells cas);
        if LENGTH=. or LENGTH=0 or LENGTH<0 then delete;* s'eliminen 2980 persones;

        LLEN=log(LENGTH+1);
        AGEXIT=AGE_RECR+LENGTH/365.25;
        PY=LENGTH/365.25;

        label   DIAGSTOM='DATA DIAGNOSTIC'
                DEATHDAY='DATA MORT'
                ENDATE='FINAL SEGUIMENT'
                LENGTH='DURADA SEGUIMENT (dies)'
                LLEN='log DURADA SEGUIMENT'
                AGEXIT='EDAT FI SEGUIMENT'
                PY='PERSON-YEARS';

        drop DATESTOM D_DEATH;
run;

* CREACIÓ DE VARIABLES D'EDAT;
data dat.PROJECTE;set dat.PROJECTE;
        kk=compress(D_BIRTH2,'/');
        kk1=input(substr(kk,1,2),2.);
        kk2=input(substr(kk,3,2),2.);
        kk3=input(substr(kk,7,2),2.);
        DATANAIX=D_BIRTH;
        if DATANAIX=. and kk1=0 and kk2=0 and kk3>0 then kk4=mdy(7,1,kk3);
        if DATANAIX=. and kk1=0 and kk2>0 and kk3>0 then kk4=mdy(kk2,15,kk3);
        if DATANAIX=. then DATANAIX=kk4;
        drop kk--kk3 kk4;
        EDAT=(D_RECRUI-DATANAIX)/365.25;
        EDAT_Q=(D_NDTQST-DATANAIX)/365.25;

        * En alguns casos hi ha informació de tabac amb la data del qüestionari no dieta en blanc. Suposo
        que l'han deguda recuperar d'algun altre lloc. Per evitar problemes amb les següents
        recodificacions assigno l'edat al reclutament quan hi ha missing. Es podrà comprovar el nombre
        de persones sense qüestionari amb la variable D_NDTQST;
        if EDAT_Q=. then EDAT_Q=EDAT;

        label    DATANAIX='Birthdate'
                 EDAT='Age (at recruitment)'
                 EDAT_Q='Age (non-dietary quest.)';

        drop D_BIRTH2 D_BIRTH;
run;

* ELIMINACIÓ DE LA INFORMACIÓ DE CÀNCER DELS FUTURS CASOS (ara controls);
data dat.PROJECTE;set dat.PROJECTE;
        if CASESTO=0 then SITESTOM='';
        if CASESTO=0 then MORPSTOM='';
        if CASESTO=0 then SDG1STOM=.;
        if CASESTO=0 then BDG1STOM='';
        if CASESTO=0 then DIAGSTOM='';
        if CASESTO=0 then BEHASTOM=.;
run;

** NIVELL ESCOLAR;
* Recupero alguna dada sobre nivell escolar. ;
data dat.PROJECTE;set dat.PROJECTE;
        if L_SCHOOL=. then L_SCHOOL=5;* Els missing els poso en "Not specified";
        if L_SCHOOL=5 and SCHOOL=0 then L_SCHOOL=0;* Els que no han anat a escola = "Sense títol";
        if L_SCHOOL=5 and A_SCHOOL<12 and A_SCHOOL^=. then L_SCHOOL=0;* Idem pels que han acabat abans dels 12;
        if L_SCHOOL=5 and A_SCHO_C<4 and A_SCHO_C^=. then L_SCHOOL=0;* Idem categòrica;
run;

** DUMMIES (variables d'ajust);
data dat.PROJECTE;set dat.PROJECTE;
        kkSEX=0;
        if SEX=2 then kkSEX=1;

        kkSCHOO1=0;
        kkSCHOO2=0;
        kkSCHOO3=0;
        kkSCHOO4=0;
        kkSCHOO5=0;
        if L_SCHOOL=1 then kkSCHOO1=1;
        if L_SCHOOL=2 then kkSCHOO2=1;
        if L_SCHOOL=3 then kkSCHOO3=1;
        if L_SCHOOL=4 then kkSCHOO4=1;
        if L_SCHOOL=5 then kkSCHOO5=1;* No hi ha missings (és la categoria 5);

        label    kkSEX='Women'
                 kkSCHOO1='Primary'
```

L

```
                kkSCHOO2='Technical'
                kkSCHOO3='Secondary'
                kkSCHOO4='University'
                kkSCHOO5='Not specified / missing';
run;


************************************************************ TABAC ****************************************;

* Descripció de les noves variables (etiquetades en català):
a)                - FUMA: CONSUM DE TABAC ACTUAL (mai, ex, actual).
                  - DURADA: TEMPS DE CONSUM DE CIGARRETES: s'haurà de calcular exclusivament com a
                    edat final (o actual) - edat inicial. Per a França s'ha usat un valor
                    aproximat de l'edat de començar a fumar (la mitjana corresponent als altres
                    centres per a la seva edat). Si només es disposa de tabac l'usarem com a proxy
                    de cigarretes.
                  - CIGDIA: QUANTITAT DE CIGARRETES/DIA: categoritzada, ja que hi ha centres
                    que no la tenen contínua. Es refereix a la darrera informació disponible.
                  - CIGPIP: CIGARS I PIPES: Crear una variable que indiqui el consum d'una o més
                    d'aquestes formes de tabac. Vist el resultat només es considera consum actual.
                  - T_NOFUMA: TEMPS EXFUMADOR: de cigarretes (o tabac si no hi ha cigarretes).
                  - EDATINI: EDAT INICI: de cigarretes (tabac si no hi ha cigarretes).
                    Categoritzada ja que a França usen aquesta classificació.

b) En cas de discrepàncies:
                   - si una variable és missing i una altra té resposta usar la que té resposta.
                   - si ambdues tenen resposta, prioritzar en l'ordre de l'apartat anterior.;


** FUMA: 0 mai fumador, 1 ex-fumador, 2 fumador actual;

data dat.PROJECTE;set dat.PROJECTE;
         FUMA=.;
         FUMA=SMOKE;
         if CIGARETT=1 or CIGARS=1 or PIPE=1 then FUMA=2;

         ** FUMA=. però variables de tabac amb resposta;

         * Si fuma algun producte actualment=fuma;
         if FUMA=. and SMOKE=. and (CIGARETT=1 or CIGARS=1 or PIPE=1) then FUMA=2;

         * Si va deixar de fumar=ex-fumador;
         if FUMA=. and SMOKE=. and (A_GIVSM>0 or A_GIVCT>0) then FUMA=1;

         * Contesta preguntes tabac actual=fuma;
         if FUMA=. and SMOKE=. and (N_CIGRET>0 or N_CIG_C>0 or N_CIG_C2>0 or F_CIGRET>0 or I_CIGRET^=.) then FUMA=2;

         ** Les següents ordres després de veure els llistats;

         * Van començar a fumar però als 40 ja no fumaven=ex-fumador;
         if FUMA=. and SMOKE=. and EDAT_Q>35 and N_CIG40=0 and (A_SMOKE>0 or A_CIGRET>0) then FUMA=1;

         * No fuma actualment i no contesta la resta=mai fuma;
         if FUMA=. and SMOKE=. and CIGARETT=0 and PIPE=0 and CIGARS=0 and A_SMOKE=. and A_SMOK_C=. and
         A_GIVSM=. and N_CIGRET=. and N_CIG_C=. and N_CIG_C2=. and F_CIGRET=. and I_CIGRET=. and A_CIGRET=.
         and N_CIG20=. and N_CIG20C=. and F_CIG20=. and N_CIG30=. and N_CIG30C=. and F_CIG30=. and
         N_CIG40=. and N_CIG40C=. and F_CIG40=. and N_CIG50=. and N_CIG50C=. and F_CIG50=. and A_GIVCT=.
         and GIV_SMOK<=0 then FUMA=0;

         * No fuma actualment (permeto fins a 2 missings entre cigarreta-cigar-pipa),
         no contesta res d'actual però va començar a fumar=ex-fumador;
         if FUMA=. and SMOKE=. and not(CIGARETT=. and CIGARS=. and PIPE=.) and
         not(CIGARETT=1 or CIGARS=1 or PIPE=1) and (A_CIGRET>0 or A_SMOKE>0) and F_CIG20=.
         and F_CIG30=. and F_CIG40=. and GIV_SMOK=. and N_CIG20<=0 and N_CIG30<=0 and N_CIG40<=0
         and N_CIG50<=0 then FUMA=1;

         * No fuma actualment (permeto fins a 2 missings entre cigarreta-cigar-pipa),
         no contesta res d'actual ni antic=mai fumador;
         if FUMA=. and SMOKE=. and not(CIGARETT=. and CIGARS=. and PIPE=.) and
         not(CIGARETT=1 or CIGARS=1 or PIPE=1) and A_CIGRET=. and A_SMOKE=. and F_CIG20=.
         and F_CIG30=. and F_CIG40=. and GIV_SMOK<=0 and N_CIG20<=0 and N_CIG30<=0 and N_CIG40<=0
         and N_CIG50<=0 then FUMA=0;

         * Casos particulars;
         if IDEPIC='92000000009121' or IDEPIC='92000000018097' then FUMA=1;
run;


** FUMADOR DE CIGARRETES:;

data dat.PROJECTE;set dat.PROJECTE;
         CIGARRET=.;
         if FUMA=0 then CIGARRET=0;* Mai fumador de cigarretes;
         if FUMA>0 and A_GIVCT>0 and CIGARETT^=1 then CIGARRET=1;* Ex fumador de cigarretes;
         if FUMA=2 and CIGARETT=1 then CIGARRET=2;* Fumador actual de cigarretes;
```

```
run;


** DURADA;

data dat.PROJECTE;set dat.PROJECTE;
         DURADA=.;
         if CIGARETT^=1 and A_CIGRET>0 and A_GIVCT>0 then DURADA=A_GIVCT-A_CIGRET;
         if CIGARETT=1 or (FUMA=2 and CIGARETT=. and DURADA=.) then DURADA=floor(EDAT_Q)-A_CIGRET;

         * A DK i S no diferenciem per tipus de tabac;
         if FUMA=2 and COUNTRY in ("8","9") then DURADA=floor(EDAT_Q)-A_SMOKE;
         if FUMA=1 and COUNTRY in ("8","9") then DURADA=A_GIVSM-A_SMOKE;

         * A França l'edat d'inici està categoritzada. He escollit els valors mitjos de la resta
         de centres per a cada franja d'edat per calcular la marca de classe;
         if DURADA=. and A_SMOK_C=1 and A_GIVSM>=14 and EDAT_Q>=14 then DURADA=A_GIVSM-14;
         if DURADA=. and A_SMOK_C=2 and A_GIVSM>=19 and EDAT_Q>=19 then DURADA=A_GIVSM-19;
         if DURADA=. and A_SMOK_C=3 and A_GIVSM>=30 and EDAT_Q>=30 then DURADA=A_GIVSM-30;
         if DURADA=. and A_SMOK_C=4 and A_GIVSM>=42 and EDAT_Q>=42 then DURADA=A_GIVSM-42;
         if FUMA=. then DURADA=.;
         if FUMA=0 then DURADA=0;
         if DURADA<0 then DURADA=.;

         * Deixa de fumar abans de la marca de classe. Escullo el punt mig de totes les possibilitats.;
         if DURADA=. and A_SMOK_C=2 and A_GIVSM=16 then DURADA=0;
         if DURADA=. and A_SMOK_C=2 and A_GIVSM in (17,18) then DURADA=1;
         if DURADA=. and A_SMOK_C=3 and A_GIVSM=26 then DURADA=0;
         if DURADA=. and A_SMOK_C=3 and A_GIVSM in (27,28) then DURADA=1;
         if DURADA=. and A_SMOK_C=3 and A_GIVSM=29 then DURADA=2;
         if DURADA=. and A_SMOK_C=4 and A_GIVSM=36 then DURADA=0;
         if DURADA=. and A_SMOK_C=4 and A_GIVSM in (37,38) then DURADA=1;
         if DURADA=. and A_SMOK_C=4 and A_GIVSM in (39,40) then DURADA=2;
         if DURADA=. and A_SMOK_C=4 and A_GIVSM=41 then DURADA=3;

         * Fumadors actuals a França;
         if DURADA=. and (CIGARETT=1 or (CIGARETT=. and FUMA=2)) and A_SMOK_C=1 then DURADA=floor(EDAT_Q)-14;
         if DURADA=. and (CIGARETT=1 or (CIGARETT=. and FUMA=2)) and A_SMOK_C=2 then DURADA=floor(EDAT_Q)-19;
         if DURADA=. and (CIGARETT=1 or (CIGARETT=. and FUMA=2)) and A_SMOK_C=3 then DURADA=floor(EDAT_Q)-30;
         if DURADA=. and (CIGARETT=1 or (CIGARETT=. and FUMA=2)) and A_SMOK_C=4 then DURADA=floor(EDAT_Q)-42;

         * Quan fa molt poc que ha començat a fumar (<1 any) pot haver-hi algun temps negatiu que el
         passo a 0;
         if DURADA=. and CIGARETT=1 and ABS(EDAT_Q-A_CIGRET)<1 and EDAT_Q>0 and A_CIGRET>0 then DURADA=0;

         * DURADA ha de ser < que EDAT;
         if DURADA>=EDAT_Q and EDAT_Q^=. then DURADA=.;

         * Si s'ha calculat la DURADA amb una data impossible cal fer-la missing. Marge d'error=1 any;
         if A_CIGRET>(EDAT_Q+1) and EDAT_Q^=. and FUMA>0 then DURADA=.;
         if A_SMOKE>(EDAT_Q+1) and EDAT_Q^=. and FUMA>0 then DURADA=.;
         if A_GIVSM>(EDAT_Q+1) and EDAT_Q^=. and FUMA>0 then DURADA=.;
         if A_GIVCT>(EDAT_Q+1) and EDAT_Q^=. and FUMA>0 then DURADA=.;

         * Cas particular: recupero 2 casos que se que fumaven abans dels 30. Hi poso la mitjana dels homes
         anglesos que pertanyen a la seva edat (+-5) i FUMA igual. En ambdós casos és 17 anys com a inici;
         if IDEPIC='41000041077683' and DURADA=. then DURADA=52;
         if IDEPIC='42000000040118' and DURADA=. then DURADA=20;
run;


** CIGDIA;

data dat.PROJECTE;set dat.PROJECTE;
         CIGDIA=.;

         * CIGDIA està codificada com 1(1-4), 2(5-14), 3(15-24) i 4(25 o més) degut a que França i Umea
         usen aquesta classificació. La conversió de les dades d'Itàlia és aproximada doncs valien
         1-3, 4-8, 9-13, 14-18, 19-23, 24-28, 29-33 i 34+.;

         CIGDIA=.;
         if (N_CIGRET<5 and N_CIGRET>0) or N_CIG_C=1 or N_CIG_C2=1 then CIGDIA=1;
         if (N_CIGRET<15 and N_CIGRET>4) or N_CIG_C=2 or N_CIG_C=3 or N_CIG_C2=2 then CIGDIA=2;
         if (N_CIGRET<25 and N_CIGRET>14) or N_CIG_C=4 or N_CIG_C=5 or N_CIG_C2=3 then CIGDIA=3;
         if N_CIGRET>24 or N_CIG_C>5 or N_CIG_C2=4 then CIGDIA=4;

         if FUMA=. then CIGDIA=.;
         if FUMA=0 then CIGDIA=0;

         * Si no tinc res a N_CIGRET (o anàleg) trec informació dels darrers períodes;
         kk20=N_CIG20C;
         kk30=N_CIG30C;
         kk40=N_CIG40C;
         kk50=N_CIG50C;
         if N_CIG20=0 and kk20=. then kk20=0;
```

```
        if N_CIG20>0 then kk20=1;
        if N_CIG20>3 then kk20=2;
        if N_CIG20>8 then kk20=3;
        if N_CIG20>13 then kk20=4;
        if N_CIG20>18 then kk20=5;
        if N_CIG20>23 then kk20=6;
        if N_CIG20>28 then kk20=7;
        if N_CIG20>33 then kk20=8;
        if N_CIG30=0 and kk30=. then kk30=0;
        if N_CIG30>0 then kk30=1;
        if N_CIG30>3 then kk30=2;
        if N_CIG30>8 then kk30=3;
        if N_CIG30>13 then kk30=4;
        if N_CIG30>18 then kk30=5;
        if N_CIG30>23 then kk30=6;
        if N_CIG30>28 then kk30=7;
        if N_CIG30>33 then kk30=8;
        if N_CIG40=0 and kk40=. then kk40=0;
        if N_CIG40>0 then kk40=1;
        if N_CIG40>3 then kk40=2;
        if N_CIG40>8 then kk40=3;
        if N_CIG40>13 then kk40=4;
        if N_CIG40>18 then kk40=5;
        if N_CIG40>23 then kk40=6;
        if N_CIG40>28 then kk40=7;
        if N_CIG40>33 then kk40=8;
        if N_CIG50=0 and kk50=. then kk50=0;
        if N_CIG50>0 then kk50=1;
        if N_CIG50>3 then kk50=2;
        if N_CIG50>8 then kk50=3;
        if N_CIG50>13 then kk50=4;
        if N_CIG50>18 then kk50=5;
        if N_CIG50>23 then kk50=6;
        if N_CIG50>28 then kk50=7;
        if N_CIG50>33 then kk50=8;

        if FUMA>0 and CIGDIA=. and kk20>0 and (EDAT_Q>19 or EDAT_Q=.) then CIGDIA2=kk20;
        if FUMA>0 and CIGDIA=. and kk30>0 and (EDAT_Q>29 or EDAT_Q=.) then CIGDIA2=kk30;
        if FUMA>0 and CIGDIA=. and kk40>0 and (EDAT_Q>39 or EDAT_Q=.) then CIGDIA2=kk40;
        if FUMA>0 and CIGDIA=. and kk50>0 and (EDAT_Q>49 or EDAT_Q=.) then CIGDIA2=kk50;

        if CIGDIA2=1 and CIGDIA=. then CIGDIA=1;
        if (CIGDIA2=2 or CIGDIA2=3) and CIGDIA=. then CIGDIA=2;
        if (CIGDIA2=4 or CIGDIA2=5) and CIGDIA=. then CIGDIA=3;
        if CIGDIA2>5 and CIGDIA=. then CIGDIA=4;

        drop KK20--KK50 CIGDIA2;
run;


** CIGPIP;

data dat.PROJECTE;set dat.PROJECTE;
        CIGPIP=.;
        if CIGARS=1 or PIPE=1 then CIGPIP=1;
        if CIGARS=0 and PIPE=0 then CIGPIP=0;
        if FUMA=. then CIGPIP=.;

        * Només es pregunta el consum actual de cigars i pipa. Per tant, no puc dir res dels ex-fumadors.
        La variable s'interpreta com fumador de pipa/cigar actual sí/no;
        if FUMA=0 or FUMA=1 then CIGPIP=0;

        * Si fuma cigarretes actualment i no diu res més interpreto que no fuma res més
        (es mantenen els missings a NL ja que ells no pregunten cigars i pipa);
        if CIGPIP=. and FUMA=2 and CIGARETT=1 and CIGARS=. and PIPE=. and COUNTRY^="5" then CIGPIP=0;
        if CIGPIP=. and FUMA=2 and CIGARETT=1 and (CIGARS=0 or PIPE=0) and not(CIGARS=1 or PIPE=1) then CIGPIP=0;

        * Si fuma actualment però no fuma cigarretes considero que ha de fumar cigars o pipes;
        if FUMA=2 and CIGARETT=0 and (PIPE=. or CIGARS=.) then CIGPIP=1;
run;


** T_NOFUMA;

data dat.PROJECTE;set dat.PROJECTE;
        T_NOFUMA=.;
        if A_GIVSM>0 then T_NOFUMA=floor(EDAT_Q-A_GIVSM);
        if A_GIVCT>0 then T_NOFUMA=floor(EDAT_Q-A_GIVCT);

        * Si fuma cigarretes actualment (o tabac a F, S, DK) temps no fuma és 0;
        if CIGARETT=1 then T_NOFUMA=0;
        if FUMA=2 and (COUNTRY='8' or COUNTRY='9' or COUNTRY='1') then T_NOFUMA=0;

        * Si fuma actualment (però no sabem què) i no hi ha res a A_GIVCT suposo que encara fuma cigarretes;
        if FUMA=2 and CIGARETT=. and A_GIVCT=. then T_NOFUMA=0;
```

```
        * Permeto un error d'un any amb les edats;
        if T_NOFUMA<0 and T_NOFUMA^=. and ABS(EDAT_Q-A_GIVCT)<1 and EDAT_Q^=. and A_GIVCT^=. then T_NOFUMA=0;
        if T_NOFUMA<0 and T_NOFUMA^=. and ABS(EDAT_Q-A_GIVSM)<1 and EDAT_Q^=. and A_GIVSM^=. then T_NOFUMA=0;

        * Temps negatius i mai fumadors;
        if T_NOFUMA<0 then T_NOFUMA=.;
        if FUMA=0 then T_NOFUMA=.;
run;


** EDATINI es categoritza, degut a França, com <16, 16-25, 26-35 i >35;

data dat.PROJECTE;set dat.PROJECTE;
        EDATINI=.;
        if (A_CIGRET<16 and A_CIGRET>0) or (A_SMOKE<16 and A_SMOKE>0) or A_SMOK_C=1 then EDATINI=1;
        if (A_CIGRET<26 and A_CIGRET>15) or (A_SMOKE<26 and A_SMOKE>15) or A_SMOK_C=2 then EDATINI=2;
        if (A_CIGRET<36 and A_CIGRET>25) or (A_SMOKE<36 and A_SMOKE>25) or A_SMOK_C=3 then EDATINI=3;
        if A_CIGRET>35 or A_SMOKE>35 or A_SMOK_C=4 then EDATINI=4;

        * Si no ha fumat o no se sap no té edat d'inici;
        if FUMA<1 then EDATINI=.;
run;


* LABELS;
data dat.PROJECTE;set dat.PROJECTE;
        label    FUMA='Smoking status'
                 CIGARRET='Cigarette smoking status'
                 DURADA='Cigarette smoking years'
                 CIGDIA='Cigarettes/day'
                 CIGPIP='Current cigar/pipe consumption'
                 T_NOFUMA='Years since give up cigarettes'
                 EDATINI='Age started smoking cigarettes';
run;

proc format library=library;
        value FUMA 0='Never' 1='Former' 2='Current';
        value CIGDIA 1='1-4' 2='5-14' 3='15-24' 4='>=25';
        value CIGPIP 0='No' 1='Yes';
        value EDATINI 1='<16' 2='16-25' 3='26-35' 4='>=36';
run;


* DUMMIES;
data dat.PROJECTE;set dat.PROJECTE;
        kkFUMA1=0;
        kkFUMA2=0;
        if FUMA=1 then kkFUMA1=1;
        if FUMA=2 then kkFUMA2=1;
        if FUMA=. then kkFUMA1=.;
        if FUMA=. then kkFUMA2=.;

        kkCIG1=0;
        kkCIG2=0;
        if CIGARRET=1 then kkCIG1=1;
        if CIGARRET=2 then kkCIG2=1;
        if CIGARRET=. then kkCIG1=.;
        if CIGARRET=. then kkCIG2=.;

        kkCIGDI1=0;
        kkCIGDI2=0;
        kkCIGDI3=0;
        kkCIGDI4=0;
        if CIGDIA=1 then kkCIGDI1=1;
        if CIGDIA=2 then kkCIGDI2=1;
        if CIGDIA=3 then kkCIGDI3=1;
        if CIGDIA=4 then kkCIGDI4=1;
        if CIGDIA=. then kkCIGDI1=.;
        if CIGDIA=. then kkCIGDI2=.;
        if CIGDIA=. then kkCIGDI3=.;
        if CIGDIA=. then kkCIGDI4=.;

        kkEDATI2=0;
        kkEDATI3=0;
        kkEDATI4=0;
        if EDATINI=2 then kkEDATI2=1;
        if EDATINI=3 then kkEDATI3=1;
        if EDATINI=4 then kkEDATI4=1;
        if EDATINI=. then kkEDATI2=.;
        if EDATINI=. then kkEDATI3=.;
        if EDATINI=. then kkEDATI4=.;

        kkCURR1=0;
        kkCURR2=0;
```

```
        kkCURR3=0;
        kkCURR4=0;
        if kkCIGDI1=1 and kkFUMA2=1 then kkCURR1=1;
        if kkCIGDI2=1 and kkFUMA2=1 then kkCURR2=1;
        if kkCIGDI3=1 and kkFUMA2=1 then kkCURR3=1;
        if kkCIGDI4=1 and kkFUMA2=1 then kkCURR4=1;
        if kkCIGDI1=. and kkFUMA2=1 then kkCURR1=.;
        if kkCIGDI1=. and kkFUMA2=1 then kkCURR2=.;
        if kkCIGDI1=. and kkFUMA2=1 then kkCURR3=.;
        if kkCIGDI1=. and kkFUMA2=1 then kkCURR4=.;

        kkTNOF2=0;
        kkTNOF3=0;
        kkTNOF4=0;
        if T_NOFUMA>=2 and T_NOFUMA<6 then kkTNOF2=1;
        if T_NOFUMA>=6 and T_NOFUMA<10 then kkTNOF3=1;
        if T_NOFUMA>=10 then kkTNOF4=1;
        if T_NOFUMA=. then kkTNOF2=.;
        if T_NOFUMA=. then kkTNOF3=.;
        if T_NOFUMA=. then kkTNOF4=.;

        kkFUMAEV=0;
        if kkFUMA1>0 or kkFUMA2>0 then kkFUMAEV=1;* kkFUMAEV 0=never 1=ever;
        if kkFUMA1=. then kkFUMAEV=.;

        kkCIG=0;
        if kkCIG1>0 or kkCIG2>0 then kkCIG=1;* kkCIG 0=never 1=ever;
        if kkCIG1=. then kkCIG=.;

        kkDUR1=0;
        kkDUR2=0;
        kkDUR3=0;
        kkDUR0=0;* la referència seran els never smokers;
        if DURADA<10 and DURADA>=0 and kkFUMAEV=1 then kkDUR0=1;
        if DURADA>=10 and DURADA<20 then kkDUR1=1;
        if DURADA>=20 and DURADA<30 then kkDUR2=1;
        if DURADA>=30 then kkDUR3=1;
        if DURADA=. then kkDUR1=.;
        if DURADA=. then kkDUR2=.;
        if DURADA=. then kkDUR3=.;
        if DURADA=. then kkDUR0=.;

        TR_DURAD=0;
        if kkDUR0=1 then TR_DURAD=1;
        if kkDUR1=1 then TR_DURAD=2;
        if kkDUR2=1 then TR_DURAD=3;
        if kkDUR3=1 then TR_DURAD=4;
        if DURADA=. then TR_DURAD=.;
run;

************************************  final part tabac ***************************;



* S'ELIMINEN ELS QUE NO TENEN INFORMACIÓ SOBRE DIETA (hi ha algun missing en QCARBO que no tindré en compte);
data dat.PROJECTE;set dat.PROJECTE;
        if QG07=. then delete;* 6334 individus eliminats;
run;

* S'ELIMINEN ELS NORUECS (ja que no tenen cap cas);
data dat.PROJECTE;set dat.PROJECTE;
        if COUNTRY='B' then delete;* 37199 individus eliminats;
run;

* Cal ordenar abans de calibrar. Com que hi ha codis IDQST cal tenir en compte el centre.;
proc sort data=dat.PROJECTE;by CENTER IDQST;run;
proc sort data=dat.R24H;by CENTER IDQST;run;



********************************** PREPARACIÓ DEL FITXER AMB ELS R24H *********************************;

* ESBORRO NORUECS;
data dat.R24H;set dat.R24H;if CENTER="B1" or CENTER="B2" then delete;run;* elimino 1798 individus;

* ESBORRO 113 INDIVIDUS QUE NO ESTAN AL FITXER PROJECTE (degut a que tenen càncer gàstric prevalent,
  o no tenen dieta baseline, o seguiment negatiu) (tots els individus figuren a ORIGINAL);
data dat.PROJECTE;set dat.PROJECTE;FITXER=1;run;* marcador de fitxer PROJECTE;
data dat.R24H;merge dat.R24H dat.PROJECTE;by CENTER IDQST;run;
data dat.R24H;set dat.R24H;
        if FITXER=. or RG07=. then delete;* 113 individus eliminats (+ els que no tenen R24H);
run;

data dat.R24H;set dat.R24H;if EXCLEIER^=2 then delete;run;* 627 individus eliminats;
```

```
data dat.R24H;set dat.R24H;keep RG07--SEASONS4;run;* elimino variables innecessàries;

* POSO LA INFORMACIÓ DEL R24H EN PROJECTE;
data dat.PROJECTE;merge dat.PROJECTE dat.R24H;by CENTER IDQST;run;


*************************************************************************************************;

* ELIMINO ELS QUE TINGUIN EXCLEIER ALT O BAIX, però abans ho descric;
proc tabulate data=dat.PROJECTE;
         class SEX;var QG07 QENER QALCOHOL;
         tables (QG07 QENER QALCOHOL),SEX*(n mean std min max);
run;

data dat.PROJECTE;set dat.PROJECTE;if EXCLEIER^=2 then delete;run;* 9413 individus eliminats;


* DESCRIPCIÓ DESPRÉS D'ELIMINAR EI/ER EXTREMS;
proc tabulate data=dat.PROJECTE;
         class SEX;var QG07 QENER QALCOHOL;
         tables (QG07 QENER QALCOHOL),SEX*(n mean std min max);
run;

* ELIMINO VARIABLES PER REDUÏR LA BASE DE DADES;
data dat.PROJECTE;set dat.PROJECTE;
         drop TRY_SMOK--N_CI50C2;
run;

data dat.PROJECTE;set dat.PROJECTE;
         length kkSEX--FITXER 3;
run;

* DESCRIPCIÓ DELS PARTICIPANTS A L'ESTUDI BASELINE;
proc tabulate data=dat.PROJECTE;
         class SEX COUNTRY;
         tables COUNTRY all, (SEX all)*(n*f=8.0);
run;

* DESCRIPCIÓ DELS PARTICIPANTS A L'ESTUDI DE CALIBRATGE;
proc tabulate data=dat.PROJECTE;
         class SEX COUNTRY;
         tables COUNTRY all, (SEX all)*(n*f=8.0);where RG07^=.;
run;


* CÀLCUL DE L'ESTACIÓ EN QUE ES VA FER EL QÜESTIONARI DE DIETA;
data dat.PROJECTE;set dat.PROJECTE;
         kk=D_DTQST;
         * 45776 individus (basicament italians) no tenen informació sobre D_DTQST (data Q dieta),
         per tant s'els hi assigna la data del Q no dieta;
         if kk=. then kk=D_NDTQST;
         if kk^=. then do;
                  MESDQ=month(kk);
                  DIADQ=day(kk);
                  ESTADQ1=0;
                  ESTADQ2=0;
                  ESTADQ3=0;
         if (MESDQ=3 and DIADQ>20) or MESDQ=4 or MESDQ=5  or (MESDQ=6 and DIADQ<21) then ESTADQ1=1;
         if (MESDQ=6 and DIADQ>20) or MESDQ=7 or MESDQ=8 or (MESDQ=9 and DIADQ<21) then ESTADQ2=1;
         if (MESDQ=9 and DIADQ>20) or MESDQ=10 or MESDQ=11 or (MESDQ=12 and DIADQ<21) then ESTADQ3=1;
         end;
         if kk=. then do;* si no tinc cap data de qüestionari (53 individus, basicament grecs) reparteixo el pes;
                  ESTADQ1=0.25;
                  ESTADQ2=0.25;
                  ESTADQ3=0.25;
         end;

         label    ESTADQ1="DQ PRIMAVERA"
                  ESTADQ2="DQ ESTIU"
                  ESTADQ3="DQ TARDOR";

         drop kk MESDQ DIADQ;
         length ESTADQ1--ESTADQ3 3;
run;


* CÀLCUL DEL PES SEGONS DIA I ESTACIÓ DEL R24H (específic per país i sexe);
proc format library=library;
         value DIAR24H 1-4="Dilluns-dijous" 5-7="Divendres-diumenge";
run;

proc tabulate data=dat.PROJECTE;
         class SEX SEASONS4 REC_DAY;format SEASONS4 seasonsb. REC_DAY diar24h.;
         tables (SEX all),SEASONS4*REC_DAY,(n*f=8.0 pctn<SEASONS4*REC_DAY>='%'*f=6.2);
         title 'Distribució general de l''estació i dia del R24H';
```

```
run;


* S'espera que cada dia de la setmana i estació de l'any estiguin igualment representat.
* Ho faig separat per sexe i país, ja que correré els models separadament. Les diferències són petites.;
proc freq data=dat.PROJECTE;
         format SEASONS4 seasonsb. REC_DAY diar24h.;
         tables COUNTRY*SEASONS4*REC_DAY/out=dat.esborra1 outpct;where SEX=1;
run;

proc freq data=dat.PROJECTE;
         format SEASONS4 seasonsb. REC_DAY diar24h.;
         tables COUNTRY*SEASONS4*REC_DAY/out=dat.esborra2 outpct;where SEX=2;
run;

data dat.esborra1;set dat.esborra1 dat.esborra2;
         if REC_DAY<5 and REC_DAY>0 then denomin=0.142857;
         if REC_DAY>4 then denomin=0.107143;
         ponderal=denomin*100/pct_tabl;
         SEX=1;
         if _n_>=74 then SEX=2;
run;

proc sort data=dat.ESBORRA1;by SEX COUNTRY SEASONS4 REC_DAY;run;


* LLISTAT DELS PESOS PER INSERTAR-LO A LA INSTRUCCIÓ DE SOTA;
proc print data=dat.esborra1;var COUNTRY SEASONS4 REC_DAY SEX PONDERAL;where SEASONS4^=.;run;

* CREACIÓ DEL PES (=variable PONDERAL d'ESBORRA1.SAS7BDAT);
data dat.PROJECTE;set dat.PROJECTE;
         if SEX=1 and COUNTRY="2" and SEASONS4=1 and REC_DAY<5 then PONDER=0.79440;
         if SEX=1 and COUNTRY="2" and SEASONS4=1 and REC_DAY>4 then PONDER=1.00615;
         if SEX=1 and COUNTRY="2" and SEASONS4=2 and REC_DAY<5 then PONDER=1.21300;
         if SEX=1 and COUNTRY="2" and SEASONS4=2 and REC_DAY>4 then PONDER=1.66955;
         if SEX=1 and COUNTRY="2" and SEASONS4=3 and REC_DAY<5 then PONDER=0.85473;
         if SEX=1 and COUNTRY="2" and SEASONS4=3 and REC_DAY>4 then PONDER=1.27671;
         if SEX=1 and COUNTRY="2" and SEASONS4=4 and REC_DAY<5 then PONDER=0.82013;
         if SEX=1 and COUNTRY="2" and SEASONS4=4 and REC_DAY>4 then PONDER=1.00615;
         if SEX=1 and COUNTRY="3" and SEASONS4=1 and REC_DAY<5 then PONDER=0.97266;
         if SEX=1 and COUNTRY="3" and SEASONS4=1 and REC_DAY>4 then PONDER=0.82269;
         if SEX=1 and COUNTRY="3" and SEASONS4=2 and REC_DAY<5 then PONDER=0.87368;
         if SEX=1 and COUNTRY="3" and SEASONS4=2 and REC_DAY>4 then PONDER=1.12500;
         if SEX=1 and COUNTRY="3" and SEASONS4=3 and REC_DAY<5 then PONDER=1.19711;
         if SEX=1 and COUNTRY="3" and SEASONS4=3 and REC_DAY>4 then PONDER=1.12500;
         if SEX=1 and COUNTRY="3" and SEASONS4=4 and REC_DAY<5 then PONDER=1.01219;
         if SEX=1 and COUNTRY="3" and SEASONS4=4 and REC_DAY>4 then PONDER=0.98810;
         if SEX=1 and COUNTRY="4" and SEASONS4=1 and REC_DAY<5 then PONDER=0.69918;
         if SEX=1 and COUNTRY="4" and SEASONS4=1 and REC_DAY>4 then PONDER=1.06933;
         if SEX=1 and COUNTRY="4" and SEASONS4=2 and REC_DAY<5 then PONDER=1.06933;
         if SEX=1 and COUNTRY="4" and SEASONS4=2 and REC_DAY>4 then PONDER=1.39835;
         if SEX=1 and COUNTRY="4" and SEASONS4=3 and REC_DAY<5 then PONDER=1.25369;
         if SEX=1 and COUNTRY="4" and SEASONS4=3 and REC_DAY>4 then PONDER=0.72714;
         if SEX=1 and COUNTRY="4" and SEASONS4=4 and REC_DAY<5 then PONDER=0.90893;
         if SEX=1 and COUNTRY="4" and SEASONS4=4 and REC_DAY>4 then PONDER=1.60399;
         if SEX=1 and COUNTRY="5" and SEASONS4=1 and REC_DAY<5 then PONDER=1.05663;
         if SEX=1 and COUNTRY="5" and SEASONS4=1 and REC_DAY>4 then PONDER=0.89286;
         if SEX=1 and COUNTRY="5" and SEASONS4=2 and REC_DAY<5 then PONDER=0.77640;
         if SEX=1 and COUNTRY="5" and SEASONS4=2 and REC_DAY>4 then PONDER=0.97758;
         if SEX=1 and COUNTRY="5" and SEASONS4=3 and REC_DAY<5 then PONDER=0.96525;
         if SEX=1 and COUNTRY="5" and SEASONS4=3 and REC_DAY>4 then PONDER=0.99947;
         if SEX=1 and COUNTRY="5" and SEASONS4=4 and REC_DAY<5 then PONDER=1.17481;
         if SEX=1 and COUNTRY="5" and SEASONS4=4 and REC_DAY>4 then PONDER=1.44009;
         if SEX=1 and COUNTRY="6" and SEASONS4=1 and REC_DAY<5 then PONDER=0.66208;
         if SEX=1 and COUNTRY="6" and SEASONS4=1 and REC_DAY>4 then PONDER=0.71858;
         if SEX=1 and COUNTRY="6" and SEASONS4=2 and REC_DAY<5 then PONDER=5.10989;
         if SEX=1 and COUNTRY="6" and SEASONS4=2 and REC_DAY>4 then PONDER=5.74863;
         if SEX=1 and COUNTRY="6" and SEASONS4=3 and REC_DAY<5 then PONDER=0.91415;
         if SEX=1 and COUNTRY="6" and SEASONS4=3 and REC_DAY>4 then PONDER=1.13231;
         if SEX=1 and COUNTRY="6" and SEASONS4=4 and REC_DAY<5 then PONDER=0.75774;
         if SEX=1 and COUNTRY="6" and SEASONS4=4 and REC_DAY>4 then PONDER=0.71858;
         if SEX=1 and COUNTRY="7" and SEASONS4=1 and REC_DAY<5 then PONDER=0.70293;
         if SEX=1 and COUNTRY="7" and SEASONS4=1 and REC_DAY>4 then PONDER=1.58160;
         if SEX=1 and COUNTRY="7" and SEASONS4=2 and REC_DAY<5 then PONDER=0.59855;
         if SEX=1 and COUNTRY="7" and SEASONS4=2 and REC_DAY>4 then PONDER=1.40483;
         if SEX=1 and COUNTRY="7" and SEASONS4=3 and REC_DAY<5 then PONDER=1.16215;
         if SEX=1 and COUNTRY="7" and SEASONS4=3 and REC_DAY>4 then PONDER=2.17111;
         if SEX=1 and COUNTRY="7" and SEASONS4=4 and REC_DAY<5 then PONDER=0.84464;
         if SEX=1 and COUNTRY="7" and SEASONS4=4 and REC_DAY>4 then PONDER=1.47421;
         if SEX=1 and COUNTRY="8" and SEASONS4=1 and REC_DAY<5 then PONDER=1.01329;
         if SEX=1 and COUNTRY="8" and SEASONS4=1 and REC_DAY>4 then PONDER=0.98364;
         if SEX=1 and COUNTRY="8" and SEASONS4=2 and REC_DAY<5 then PONDER=0.94720;
         if SEX=1 and COUNTRY="8" and SEASONS4=2 and REC_DAY>4 then PONDER=1.29563;
         if SEX=1 and COUNTRY="8" and SEASONS4=3 and REC_DAY<5 then PONDER=1.36161;
         if SEX=1 and COUNTRY="8" and SEASONS4=3 and REC_DAY>4 then PONDER=1.35533;
```

```
        if SEX=1 and COUNTRY="8" and SEASONS4=4 and REC_DAY<5 then PONDER=0.69406;
        if SEX=1 and COUNTRY="8" and SEASONS4=4 and REC_DAY>4 then PONDER=0.84514;
        if SEX=1 and COUNTRY="9" and SEASONS4=1 and REC_DAY<5 then PONDER=0.80543;
        if SEX=1 and COUNTRY="9" and SEASONS4=1 and REC_DAY>4 then PONDER=2.18894;
        if SEX=1 and COUNTRY="9" and SEASONS4=2 and REC_DAY<5 then PONDER=1.87192;
        if SEX=1 and COUNTRY="9" and SEASONS4=2 and REC_DAY>4 then PONDER=2.75097;
        if SEX=1 and COUNTRY="9" and SEASONS4=3 and REC_DAY<5 then PONDER=0.83260;
        if SEX=1 and COUNTRY="9" and SEASONS4=3 and REC_DAY>4 then PONDER=1.14366;
        if SEX=1 and COUNTRY="9" and SEASONS4=4 and REC_DAY<5 then PONDER=0.48469;
        if SEX=1 and COUNTRY="9" and SEASONS4=4 and REC_DAY>4 then PONDER=1.08862;
        if SEX=2 and COUNTRY="1" and SEASONS4=1 and REC_DAY<5 then PONDER=0.59823;
        if SEX=2 and COUNTRY="1" and SEASONS4=1 and REC_DAY>4 then PONDER=1.01179;
        if SEX=2 and COUNTRY="1" and SEASONS4=2 and REC_DAY<5 then PONDER=1.51806;
        if SEX=2 and COUNTRY="1" and SEASONS4=2 and REC_DAY>4 then PONDER=2.92294;
        if SEX=2 and COUNTRY="1" and SEASONS4=3 and REC_DAY<5 then PONDER=0.91542;
        if SEX=2 and COUNTRY="1" and SEASONS4=3 and REC_DAY>4 then PONDER=1.51004;
        if SEX=2 and COUNTRY="1" and SEASONS4=4 and REC_DAY<5 then PONDER=0.69347;
        if SEX=2 and COUNTRY="1" and SEASONS4=4 and REC_DAY>4 then PONDER=1.17054;
        if SEX=2 and COUNTRY="2" and SEASONS4=1 and REC_DAY<5 then PONDER=0.66949;
        if SEX=2 and COUNTRY="2" and SEASONS4=1 and REC_DAY>4 then PONDER=1.05043;
        if SEX=2 and COUNTRY="2" and SEASONS4=2 and REC_DAY<5 then PONDER=1.08403;
        if SEX=2 and COUNTRY="2" and SEASONS4=2 and REC_DAY>4 then PONDER=2.15252;
        if SEX=2 and COUNTRY="2" and SEASONS4=3 and REC_DAY<5 then PONDER=0.91660;
        if SEX=2 and COUNTRY="2" and SEASONS4=3 and REC_DAY>4 then PONDER=1.08966;
        if SEX=2 and COUNTRY="2" and SEASONS4=4 and REC_DAY<5 then PONDER=0.99472;
        if SEX=2 and COUNTRY="2" and SEASONS4=4 and REC_DAY>4 then PONDER=1.01786;
        if SEX=2 and COUNTRY="3" and SEASONS4=1 and REC_DAY<5 then PONDER=0.79447;
        if SEX=2 and COUNTRY="3" and SEASONS4=1 and REC_DAY>4 then PONDER=0.97890;
        if SEX=2 and COUNTRY="3" and SEASONS4=2 and REC_DAY<5 then PONDER=0.74444;
        if SEX=2 and COUNTRY="3" and SEASONS4=2 and REC_DAY>4 then PONDER=1.32237;
        if SEX=2 and COUNTRY="3" and SEASONS4=3 and REC_DAY<5 then PONDER=1.15517;
        if SEX=2 and COUNTRY="3" and SEASONS4=3 and REC_DAY>4 then PONDER=1.46359;
        if SEX=2 and COUNTRY="3" and SEASONS4=4 and REC_DAY<5 then PONDER=0.92627;
        if SEX=2 and COUNTRY="3" and SEASONS4=4 and REC_DAY>4 then PONDER=1.23566;
        if SEX=2 and COUNTRY="4" and SEASONS4=1 and REC_DAY<5 then PONDER=0.66840;
        if SEX=2 and COUNTRY="4" and SEASONS4=1 and REC_DAY>4 then PONDER=1.00871;
        if SEX=2 and COUNTRY="4" and SEASONS4=2 and REC_DAY<5 then PONDER=1.04043;
        if SEX=2 and COUNTRY="4" and SEASONS4=2 and REC_DAY>4 then PONDER=1.42611;
        if SEX=2 and COUNTRY="4" and SEASONS4=3 and REC_DAY<5 then PONDER=1.06044;
        if SEX=2 and COUNTRY="4" and SEASONS4=3 and REC_DAY>4 then PONDER=0.99656;
        if SEX=2 and COUNTRY="4" and SEASONS4=4 and REC_DAY<5 then PONDER=0.85493;
        if SEX=2 and COUNTRY="4" and SEASONS4=4 and REC_DAY>4 then PONDER=1.83810;
        if SEX=2 and COUNTRY="5" and SEASONS4=1 and REC_DAY<5 then PONDER=1.01542;
        if SEX=2 and COUNTRY="5" and SEASONS4=1 and REC_DAY>4 then PONDER=1.09762;
        if SEX=2 and COUNTRY="5" and SEASONS4=2 and REC_DAY<5 then PONDER=0.83063;
        if SEX=2 and COUNTRY="5" and SEASONS4=2 and REC_DAY>4 then PONDER=0.81931;
        if SEX=2 and COUNTRY="5" and SEASONS4=3 and REC_DAY<5 then PONDER=1.02902;
        if SEX=2 and COUNTRY="5" and SEASONS4=3 and REC_DAY>4 then PONDER=1.20052;
        if SEX=2 and COUNTRY="5" and SEASONS4=4 and REC_DAY<5 then PONDER=1.08470;
        if SEX=2 and COUNTRY="5" and SEASONS4=4 and REC_DAY>4 then PONDER=1.08047;
        if SEX=2 and COUNTRY="6" and SEASONS4=1 and REC_DAY<5 then PONDER=0.75120;
        if SEX=2 and COUNTRY="6" and SEASONS4=1 and REC_DAY>4 then PONDER=0.62161;
        if SEX=2 and COUNTRY="6" and SEASONS4=2 and REC_DAY<5 then PONDER=2.18012;
        if SEX=2 and COUNTRY="6" and SEASONS4=2 and REC_DAY>4 then PONDER=5.37246;
        if SEX=2 and COUNTRY="6" and SEASONS4=3 and REC_DAY<5 then PONDER=0.76263;
        if SEX=2 and COUNTRY="6" and SEASONS4=3 and REC_DAY>4 then PONDER=1.40588;
        if SEX=2 and COUNTRY="6" and SEASONS4=4 and REC_DAY<5 then PONDER=0.88748;
        if SEX=2 and COUNTRY="6" and SEASONS4=4 and REC_DAY>4 then PONDER=0.84038;
        if SEX=2 and COUNTRY="7" and SEASONS4=1 and REC_DAY<5 then PONDER=0.59566;
        if SEX=2 and COUNTRY="7" and SEASONS4=1 and REC_DAY>4 then PONDER=1.27966;
        if SEX=2 and COUNTRY="7" and SEASONS4=2 and REC_DAY<5 then PONDER=0.54025;
        if SEX=2 and COUNTRY="7" and SEASONS4=2 and REC_DAY>4 then PONDER=1.26536;
        if SEX=2 and COUNTRY="7" and SEASONS4=3 and REC_DAY<5 then PONDER=1.50249;
        if SEX=2 and COUNTRY="7" and SEASONS4=3 and REC_DAY>4 then PONDER=1.96957;
        if SEX=2 and COUNTRY="7" and SEASONS4=4 and REC_DAY<5 then PONDER=1.20319;
        if SEX=2 and COUNTRY="7" and SEASONS4=4 and REC_DAY>4 then PONDER=1.81200;
        if SEX=2 and COUNTRY="8" and SEASONS4=1 and REC_DAY<5 then PONDER=0.86832;
        if SEX=2 and COUNTRY="8" and SEASONS4=1 and REC_DAY>4 then PONDER=0.91789;
        if SEX=2 and COUNTRY="8" and SEASONS4=2 and REC_DAY<5 then PONDER=1.02032;
        if SEX=2 and COUNTRY="8" and SEASONS4=2 and REC_DAY>4 then PONDER=1.01074;
        if SEX=2 and COUNTRY="8" and SEASONS4=3 and REC_DAY<5 then PONDER=1.27053;
        if SEX=2 and COUNTRY="8" and SEASONS4=3 and REC_DAY>4 then PONDER=1.24898;
        if SEX=2 and COUNTRY="8" and SEASONS4=4 and REC_DAY<5 then PONDER=0.82383;
        if SEX=2 and COUNTRY="8" and SEASONS4=4 and REC_DAY>4 then PONDER=1.05974;
        if SEX=2 and COUNTRY="9" and SEASONS4=1 and REC_DAY<5 then PONDER=0.78114;
        if SEX=2 and COUNTRY="9" and SEASONS4=1 and REC_DAY>4 then PONDER=2.21391;
        if SEX=2 and COUNTRY="9" and SEASONS4=2 and REC_DAY<5 then PONDER=1.60245;
        if SEX=2 and COUNTRY="9" and SEASONS4=2 and REC_DAY>4 then PONDER=2.19085;
        if SEX=2 and COUNTRY="9" and SEASONS4=3 and REC_DAY<5 then PONDER=0.98052;
        if SEX=2 and COUNTRY="9" and SEASONS4=3 and REC_DAY>4 then PONDER=1.33963;
        if SEX=2 and COUNTRY="9" and SEASONS4=4 and REC_DAY<5 then PONDER=0.46352;
        if SEX=2 and COUNTRY="9" and SEASONS4=4 and REC_DAY>4 then PONDER=1.10696;
        label PONDER="PONDERALS SEGONS DIA I ESTACIÓ R24H PER PAÍS I SEXE";
run;
```

```
* Tot i haver-hi alguns ponderals una mica extrems, mín=0.46 i màx=5.75, els factors de calibratge
són prou semblants ponderant o no.;
proc freq data=dat.projecte;tables ponder;run;

* Cal crear DUMMIES per centre i interacció país*aliment si volem usar el proc REG;
data dat.PROJECTE;set dat.PROJECTE;
        * dummies centre;
        c11=0;
        c12=0;
        c13=0;
        c14=0;
        c21=0;
        c22=0;
        c23=0;
        c24=0;
        c25=0;
        c31=0;
        c32=0;
        c33=0;
        c34=0;
        c35=0;
        c41=0;
        c42=0;
        c43=0;
        c51=0;
        c52=0;
        c61=0;
        c71=0;
        c72=0;
        c81=0;
        c82=0;
        c91=0;
        c92=0;
        if CNTR_C="11" then c11=1;
        if CNTR_C="12" then c12=1;
        if CNTR_C="13" then c13=1;
        if CNTR_C="14" then c14=1;
        if CNTR_C="21" then c21=1;
        if CNTR_C="22" then c22=1;
        if CNTR_C="23" then c23=1;
        if CNTR_C="24" then c24=1;
        if CNTR_C="25" then c25=1;
        if CNTR_C="31" then c31=1;
        if CNTR_C="32" then c32=1;
        if CNTR_C="33" then c33=1;
        if CNTR_C="34" then c34=1;
        if CNTR_C="35" then c35=1;
        if CNTR_C="41" then c41=1;
        if CNTR_C="42" then c42=1;
        if CNTR_C="43" then c43=1;
        if CNTR_C="51" then c51=1;
        if CNTR_C="52" then c52=1;
        if CNTR_C="61" then c61=1;
        if CNTR_C="71" then c71=1;
        if CNTR_C="72" then c72=1;
        if CNTR_C="81" then c81=1;
        if CNTR_C="82" then c82=1;
        if CNTR_C="91" then c91=1;
        if CNTR_C="92" then c92=1;

        * dummies interacció aliment i país;
        c07_1=0;
        c07_2=0;
        c07_3=0;
        c07_4=0;
        c07_5=0;
        c07_6=0;
        c07_7=0;
        c07_8=0;
        c07_9=0;
        if COUNTRY="1" then c07_1=QG07;
        if COUNTRY="2" then c07_2=QG07;
        if COUNTRY="3" then c07_3=QG07;
        if COUNTRY="4" then c07_4=QG07;
        if COUNTRY="5" then c07_5=QG07;
        if COUNTRY="6" then c07_6=QG07;
        if COUNTRY="7" then c07_7=QG07;
        if COUNTRY="8" then c07_8=QG07;
        if COUNTRY="9" then c07_9=QG07;

        length c11--c92 3;
run;
```

```
* DETECCIÓ D'OUTLIERS I DEFINICIÓ DE NO CONSUMIDORS;
proc univariate data=dat.PROJECTE;var QG07;where QG07>0 and SEX=1;run;

proc means data=dat.PROJECTE n nmiss mean std min max;var QG07 RG07;where SEX=1;run;

proc univariate data=dat.PROJECTE;var QG07;where QG07>0 and SEX=2;run;

proc means data=dat.PROJECTE n nmiss mean std min max;var QG07 RG07;where SEX=2;run;

* Com a mesura de control, no inclouré en la calibratge els centres amb (mitjana Q)/(mitjana R) <0.5 o >2.0
  ja que indicaria que no reporten el mateix tipus d'aliment;

options pagesize=500 linesize=140 nonumber nodate;run;

proc tabulate data=dat.PROJECTE;
         class SEX CNTR_C;var QG07 RG07;
         tables CNTR_C,(QG07 RG07)*(mean);
         where RG07^=. and SEX=1;
run;

proc tabulate data=dat.PROJECTE;
         class SEX CNTR_C;var QG07 RG07;
         tables CNTR_C,(QG07 RG07)*(mean);
         where RG07^=. and SEX=2;
run;

* No cal excloure cap centre finalment;

options pagesize=500 linesize=100 nonumber nodate;run;

*** CALIBPROJ.SAS crea les variables calibrades per PROJECTE ***

Córrer aquest programa (o part d'ell) per obtenir els valors calibrats;
%include "C:\GP\Metodologia\Calibratge\Projecte UPC\prg\CALIBPROJ.SAS";



*************************************************************************************************
*                                          DESCRIPTIUS                                          *
*************************************************************************************************;


* COMPARACIÓ DE LES MESURES BASELINE, R24H I CALIBRADES PER PAÍS I SEXE, en el total de la mostra i en els
  individus de l'estudi de calibratge;
proc tabulate data=dat.PROJECTE;
         class SEX COUNTRY;var QG07 RG07 CQG07;
         tables SEX,COUNTRY*(QG07 RG07 CQG07),(n*f=8.0 mean std min*f=8.0 max*f=8.0)/rts=30;
run;

proc tabulate data=dat.PROJECTE;
         class SEX COUNTRY;var QG07 RG07 CQG07;
         tables SEX*COUNTRY*(QG07 RG07 CQG07),(n*f=8.0 mean std min*f=8.0 max*f=8.0)/rts=30;
         where RG07^=.;
run;


* CÀLCUL DE CATEGORIES DE CONSUM;
proc tabulate data=dat.PROJECTE;
         var QG07;class SEX;
         table (QG07)*(P25 P50 P75),SEX;
         title 'Llistat dels punts de tall';
run;

data dat.homes;set dat.PROJECTE;
         if sex=2 then delete;
run;

data dat.dones;set dat.PROJECTE;
         if sex=1 then delete;
run;

proc rank data=dat.HOMES out=dat.HOMES groups=4;
         var QG07;
         ranks tr07;
run;

proc rank data=dat.DONES out=dat.DONES groups=4;
         var QG07;
         ranks tr07;
run;

data dat.PROJECTE;set dat.HOMES dat.DONES;run;

* QUARTILS DE LA VARIABLE ORIGINAL PER PAÍS;
```

```
proc tabulate data=dat.PROJECTE;
         class tr07 COUNTRY SEX;
         tables SEX,COUNTRY, tr07*(pctn <tr07>='%'*f=5.1);
run;

* QUARTILS USANT LA VARIABLE CALIBRADA;
data dat.PROJECTE;set dat.PROJECTE;
         if CQG07<78.3 and CQG07^=. and SEX=1 then ctr07=0;
         if CQG07>=78.3 and CQG07<119 and SEX=1 then ctr07=1;
         if CQG07>=119 and CQG07<167 and SEX=1 then ctr07=2;
         if CQG07>=167 and SEX=1 then ctr07=3;
         if CQG07<53.0 and CQG07^=. and SEX=2 then ctr07=0;
         if CQG07>=53.0 and CQG07<86.0 and SEX=2 then ctr07=1;
         if CQG07>=86.0 and CQG07<121 and SEX=2 then ctr07=2;
         if CQG07>=121 and SEX=2 then ctr07=3;

         label ctr07="CARN calibrada en categories FFQ";
         length ctr07 3;
run;

proc tabulate data=dat.PROJECTE;
         class ctr07 COUNTRY SEX;
         tables SEX,COUNTRY, ctr07*(pctn <ctr07>='%'*f=5.1);
run;


* CÀLCUL DELS QUARTILS DE LA VARIABLE CALIBRADA;
data dat.homes;set dat.PROJECTE;
         if sex=2 then delete;
run;

data dat.dones;set dat.PROJECTE;
         if sex=1 then delete;
run;

proc rank data=dat.HOMES out=dat.HOMES groups=4;
         var CQG07;
         ranks ktr07;
run;

proc rank data=dat.DONES out=dat.DONES groups=4;
         var CQG07;
         ranks ktr07;
run;

data dat.PROJECTE;set dat.HOMES dat.DONES;
         label ktr07="CARN calibrada. Quartils";
         length ktr07 3;
run;

proc tabulate data=dat.PROJECTE;
         class ktr07 COUNTRY SEX;
         tables SEX,(COUNTRY all), ktr07*(pctn <ktr07>='%'*f=5.1);
run;



*************************************************************************************************
*                                    AJUST DEL MODEL DE CALIBRATGE                              *
*************************************************************************************************;

proc tabulate data=dat.PROJECTE;
         class COUNTRY SEX;var stude07;
         tables SEX,(COUNTRY all)*(stude07),(n mean std min max);
         title 'Residus estudentitzats';
run;

proc tabulate data=dat.PROJECTE;
         class COUNTRY SEX;var press07;
         tables SEX,(COUNTRY all)*(press07),(n mean std min max);
         title 'Residus sense l''observació i-èssima';
run;

proc tabulate data=dat.PROJECTE;
         class COUNTRY SEX;var dfits07;
         tables SEX,(COUNTRY all)*(dfits07),(n mean std min max);
         title 'Influència del valor predit';
run;

proc tabulate data=dat.PROJECTE;
         class COUNTRY SEX;var cook07;
         tables SEX,(COUNTRY all)*(cook07),(n mean std min max);
         title 'Cook';
run;
```

```
proc tabulate data=dat.PROJECTE;
        class COUNTRY SEX;var h07;
        tables SEX,(COUNTRY all)*(h07),(n mean std min max);
        title 'Leverage';
run;


* TRANSFORMACIÓ DE LA VARIABLE CARN PER INTENTAR UN MILLOR AJUST;

data dat.PROJECTE;set dat.PROJECTE;
        ARQ07=sqrt(QG07);
        ARR07=sqrt(RG07);
        LGQ07=log(QG07+0.1)/log(2);
        LGR07=log(RG07+0.1)/log(2);

        label    ARQ07='Arrel quadrada CARN (Q)'
                 ARR07='Arrel quadrada CARN (R)'
                 LGQ07='Log-2 CARN (Q)'
                 LGR07='Log-2 CARN (R)';
run;



*************************************************************************************************************
*                                          MODEL DE MALALTIA                                                *
*************************************************************************************************************;

* CREO UNA BD REDUÏDA PER TREBALLAR MÉS DEPRESSA;
data dat.RAPID;set dat.PROJECTE;
        keep IDEPIC COUNTRY CNTR_C SEX QG07 RG07 kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4
        kkSCHOO5 QENER LENGTH AGE_RECR AGEXIT CASESTO DEATHDAY CQG07 tr07 ctr07 ktr07 ARQ07--CAR07 CLG07
        COUT07 CNC07 CEN07 c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61
        c71 c72 c81 c82 c91 c92;
run;

data dat.RAPID;set dat.RAPID;
        * Re-escalo les variables a centenes de gram per obtenir estimadors més grans;
        CQG07c=CQG07/100;
        QG07c=QG07/100;

        * Cal crear les interaccions a mà!;
        SEXQ07=SEX*QG07c;
        SEXC07=SEX*CQG07c;

        * Cal crear una variable indicadora de no consumidors;
        z07=0;
        if QG07=0 then z07=1;
run;

proc sort data=dat.RAPID;by SEX;run;

** NOTA sobre l'estratificació: és preferible estratificar per centre (CNTR_C). El problema és que si algun
centre no té cap cas queda exclós del model. Així, és preferible estratificar per país (COUNTRY). S'ajustarà
per centre (a part d'estratificar per país). Els resultats són bastant semblants si ajustem per centre, o
estratifiquem per centre o estratifiquem per país i ajustem per centre alhora. Varien bastant (p.ex. HR de
1.51 a 1.97) si no tenim en compte el centre i només estratifiquem per país;


* 1.- S'escullen (basats en altres estudis) les variables d'ajust TABAC, BMI, ESCOLARITAT i ENERGIA.
Per a cada sexe miro si puc eliminar alguna d'aquestes variables;
proc phreg data=dat.RAPID;
        model (AGE_RECR,AGEXIT)*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
                                            QENER CQG07c c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34
                                            c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits;
        TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        by SEX;
        title 'CARN. VARIABLE CALIBRADA (CONTÍNUA).';
run;


* 2.- Per estar segur d'eliminar variables miro que passa sense calibrar;
proc phreg data=dat.RAPID;
        model (AGE_RECR,AGEXIT)*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                            QG07c c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                            c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits;
        TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        by SEX;
        title 'CARN. VARIABLE ORIGINAL (CONTÍNUA).';
run;


* 3.- Repeteixo els 2 models anteriors sense energia;
proc phreg data=dat.RAPID;* Calibrat;
```

```
        model (AGE_RECR,AGEXIT)*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
                                            CQG07c z07 c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35
                                            c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits;
        TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        by SEX;
        title 'CARN. VARIABLE CALIBRADA (CONTÍNUA).';
run;

proc phreg data=dat.RAPID;* Original;
        model (AGE_RECR,AGEXIT)*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
                                            QG07c c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41
                                            c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits;
        TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        by SEX;
        title 'CARN. VARIABLE ORIGINAL (CONTÍNUA).';
run;


* 4.- Repeteixo els 2 models anteriors sense nivell educatiu;
proc phreg data=dat.RAPID;* Calibrat;
        model (AGE_RECR,AGEXIT)*CASESTO(0)= kkFUMA1 kkFUMA2 BMI QENER CQG07c z07 c11 c12 c13 c14 c21 c22 c23 c24
                                            c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91
                                            c92/risklimits;
        TEST kkFUMA1, kkFUMA2;
        strata COUNTRY;
        by SEX;
        title 'CARN. VARIABLE CALIBRADA (CONTÍNUA).';
run;

proc phreg data=dat.RAPID;* Original;
        model (AGE_RECR,AGEXIT)*CASESTO(0)= kkFUMA1 kkFUMA2 BMI QENER QG07c c11 c12 c13 c14 c21 c22 c23 c24
                                            c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91
                                            c92/risklimits;
        TEST kkFUMA1, kkFUMA2;
        strata COUNTRY;
        by SEX;
        title 'CARN. VARIABLE ORIGINAL (CONTÍNUA).';
run;


* 5.- Miro la relació entre energia i nivell educatiu i el CG i el consum de carn;
proc phreg data=dat.RAPID;* Energia;
        model (AGE_RECR,AGEXIT)*CASESTO(0)= QENER c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                            c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits;
        strata COUNTRY;
        by SEX;
        title 'CARN. VARIABLE CALIBRADA (CONTÍNUA).';
run;

proc phreg data=dat.RAPID;* Nivell educatiu;
        model (AGE_RECR,AGEXIT)*CASESTO(0)= kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 c11 c12 c13 c14 c21 c22
                                            c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81
                                            c82 c91 c92/risklimits;
        TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        by SEX;
        title 'CARN. VARIABLE CALIBRADA (CONTÍNUA).';
run;

proc corr data=dat.projecte;var CQG07 QENER;by sex;run;

proc tabulate data=dat.projecte;var CQG07;
        class L_SCHOOL SEX;
        tables SEX, L_SCHOOL,CQG07*mean;
run;


* 6.- Cal separar per sexe?;
proc phreg data=dat.RAPID;
        model (AGE_RECR,AGEXIT)*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                            CQG07c z07 SEX SEXC07 c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33
                                            c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits;
        strata COUNTRY;
        title 'CARN. VARIABLE CALIBRADA (CONTÍNUA).';
run;


*** Model escollit ***;

* Nota: aquest model és equivalent a el que posa (AGE_RECR,AGEXIT) com a escala de temps,
però cal especificar-lo així per a poder calcular els residus i altres indicadors;
proc phreg data=dat.RAPID;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
```

```
                                    CQG07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41
                                   c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
        TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        id IDEPIC LENGTH CNTR_C;
        output out=dat.ajCox ld=influ1 xbeta=xb lmax=influ2 resdev=resdev07 resmart=mart07
        ressch=pha1 pha2 pha3 pha4 pha5 pha6 pha7 pha8 pha9 pha10 pha11 pha12;
        title 'CARN. VARIABLE CALIBRADA (CONTÍNUA).';
run;

* CREACIÓ D'UNA BD AMB EL 10% DELS CONTROLS I TOTS ELS CASOS PER FER ALGUNS GRÀFICS;
data dat.ajCox;set dat.ajCox;ALEATORI=ranuni(1969);run;
data dat.ajcox10;set dat.ajcox;if casesto=0 and ALEATORI>0.1 then delete;run;


* MODEL ESCOLLIT PERÒ USANT LA VARIABLE NO CALIBRADA;
proc phreg data=dat.RAPID;
        model (AGE_RECR,AGEXIT)*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                            QG07c SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41
                                            c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits;
        TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        title 'CARN. VARIABLE ORIGINAL (CONTÍNUA).';
run;


* REPETICIÓ DEL MODEL EXCLOENT PUNTS INFLUENTS;
proc phreg data=dat.AJCOX;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                 CQG07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41
                                 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
        TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        where INFLU1<=0.8;
        title 'CARN. VARIABLE CALIBRADA (CONTÍNUA).';
run;


* REPETICIÓ DEL MODEL ESTRATIFICANT PER EDAT;
proc format library=library;value edat 0-45='1' 45-55='2' 55-65='3' 65-120='4';run;

proc phreg data=dat.RAPID;
        format age_recr edat.;
        model (AGE_RECR,AGEXIT)*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                            CQG07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35
                                            c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits;
        strata COUNTRY AGE_RECR;
run;


* REPETICIÓ DEL MODEL EXCLOENT ELS SEGUIT MENYS DE 2 ANYS;
proc phreg data=dat.RAPID;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                  CQG07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
        TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        where LENGTH>730;
run;

proc format library=library;value llarg 0-730='=<2 anys' 731-10000='>2 anys';
proc ttest data=dat.RAPID;format LENGTH llarg.;class LENGTH;var QG07;where CASESTO=1;run;


* REPETICIÓ DEL MODEL PAÍS PER PAÍS;
proc sort data=dat.RAPID;by IDEPIC;run;

proc phreg data=dat.RAPID;* Calibrat;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                 CQG07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
        by COUNTRY;
run;

proc phreg data=dat.RAPID;* Original;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                 QG07c SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
        by COUNTRY;
run;


data dat.RAPID;set dat.RAPID;* Càlcul de la interacció;
        c07_1=0;
        c07_2=0;
```

```
        c07_3=0;
        c07_4=0;
        c07_5=0;
        c07_6=0;
        c07_7=0;
        c07_8=0;
        c07_9=0;
        if COUNTRY="1" then c07_1=QG07c;
        if COUNTRY="2" then c07_2=QG07c;
        if COUNTRY="3" then c07_3=QG07c;
        if COUNTRY="4" then c07_4=QG07c;
        if COUNTRY="5" then c07_5=QG07c;
        if COUNTRY="6" then c07_6=QG07c;
        if COUNTRY="7" then c07_7=QG07c;
        if COUNTRY="8" then c07_8=QG07c;
        if COUNTRY="9" then c07_9=QG07c;

        cc07_1=0;
        cc07_2=0;
        cc07_3=0;
        cc07_4=0;
        cc07_5=0;
        cc07_6=0;
        cc07_7=0;
        cc07_8=0;
        cc07_9=0;
        if COUNTRY="1" then cc07_1=CQG07c;
        if COUNTRY="2" then cc07_2=CQG07c;
        if COUNTRY="3" then cc07_3=CQG07c;
        if COUNTRY="4" then cc07_4=CQG07c;
        if COUNTRY="5" then cc07_5=CQG07c;
        if COUNTRY="6" then cc07_6=CQG07c;
        if COUNTRY="7" then cc07_7=CQG07c;
        if COUNTRY="8" then cc07_8=CQG07c;
        if COUNTRY="9" then cc07_9=CQG07c;
run;

proc phreg data=dat.RAPID;* Model amb interacció calibrat;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                 CQG07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92 cc07_1 cc07_2 cc07_3 cc07_4 cc07_5
                                 cc07_6 cc07_7 cc07_8 cc07_9/risklimits entrytime=AGE_RECR;
run;

proc phreg data=dat.RAPID;* Model amb interacció original;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                 QG07c SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92 c07_1 c07_2 c07_3 c07_4 c07_5 c07_6
                                 c07_7 c07_8 c07_9/risklimits entrytime=AGE_RECR;
run;

* COMPARACIÓ DELS MODELS AMB I SENSE INTERACCIÓ PAÍS-CARN PER VEURE HOMOGENEÏTAT;

proc phreg data=dat.kk;* Calibrat;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                 CQG07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
run;

proc phreg data=dat.RAPID;* Original;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                 QG07c SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
run;

*************************************************************************************************************
*                              MODEL DE MALALTIA AMB TRANSFORMACIONS O EXCLUSIONS                           *
*************************************************************************************************************;

* Re-escalat a 100 g/dia;
data dat.RAPID;set dat.RAPID;
        COUT07c=COUT07/100;
        Cnc07c=Cnc07/100;
        Cen07c=Cen07/100;
run;

* TRANSFORMACIÓ ARREL QUADRADA;
proc phreg data=dat.RAPID;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                 CAR07 z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41
                                 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
        TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        title 'CARN. VARIABLE CALIBRADA (arrel).';
run;
```

```
* TRANSFORMACIÓ LOG-2;
proc phreg data=dat.RAPID;
         model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                  CLG07 z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41
                                  c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
         TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
         strata COUNTRY;
         title 'CARN. VARIABLE CALIBRADA (log-2).';
run;


* EXCLOENT >P99;
proc phreg data=dat.RAPID;
         model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                  COUT07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41
                                  c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
         TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
         strata COUNTRY;
         title 'CARN. VARIABLE CALIBRADA (truncada).';
run;


* FUSIONANT <3 GRAMS AMB NO CONSUMIDORS;
proc phreg data=dat.RAPID;
         model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                  Cnc07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41
                                  c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
         TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
         strata COUNTRY;
         title 'CARN. VARIABLE CALIBRADA (<3=no consumidor).';
run;


* DESPRÉS D'AJUSTAR PER ENERGIA EN EL MODEL DE CALIBRATGE;
proc phreg data=dat.RAPID;
         model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                  Cen07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41
                                  c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
         TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
         strata COUNTRY;
         title 'CARN. VARIABLE CALIBRADA (ajustada per energia al calibrar).';
run;


* REPETICIÓ DEL MODEL USANT CATEGORIES, PER VEURE SI EXISTEIX CERTA LINEALITAT.
NOTA: L'ús de categories de la variable calibrada no queda gaire demostrat en la literatura. Els resultats
obtinguts podrien ser espuris;

data dat.rapid;set dat.rapid;
         ktr07_2=0;
         ktr07_3=0;
         ktr07_4=0;
         if ktr07=1 then ktr07_2=1;
         if ktr07=2 then ktr07_3=1;
         if ktr07=3 then ktr07_4=1;
run;

proc phreg data=dat.RAPID;* Quartils;
         model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER ktr07_2
                                  ktr07_3 ktr07_4 z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35
                                  c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
         TEST ktr07_2, ktr07_3, ktr07_4;
         strata COUNTRY;
         title 'CARN. VARIABLE EN QUARTILS CALIBRADA.';
run;


proc phreg data=dat.RAPID;* Tendència;
         model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER ktr07 z07
                                  SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52
                                  c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
         strata COUNTRY;
         title 'CARN. VARIABLE TENDÈNCIA CALIBRADA.';
run;


* NOTA: No té sentit fer l'anàlisi per país amb variables categòriques. Massa inestabilitat dels coeficients;



* EXCLUSIÓ DELS MORTS PER CAUSA DIFERENT DE CG PER VEURE SI HI HA DIFERÈNCIES;
proc phreg data=dat.RAPID;
         model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
```

```
                                    cQG07 z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                    c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
        strata COUNTRY;
        where DEATHDAY=. or CASESTO=1;
        title 'CARN. VARIABLE CALIBRADA (contínua).';
run;


* REPETICIÓ DEL MODEL EXCLOENT ELS QUE VAN PARTICIPAR A L'ESTUDI DE CALIBRATGE;
proc phreg data=dat.RAPID;
        model AGEXIT*CASESTO(0)= kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5 QENER
                                    CQG07c z07 SEX c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42
                                    c43 c51 c52 c61 c71 c72 c81 c82 c91 c92/risklimits entrytime=AGE_RECR;
        TEST kkFUMA1, kkFUMA2;TEST kkSCHOO1, kkSCHOO2, kkSCHOO3, kkSCHOO4, kkSCHOO5;
        strata COUNTRY;
        where RG07=.;
        title 'CARN. VARIABLE CALIBRADA (CONTÍNUA).';
run;



*************************************************************************************************************
*                                        MÉS TAULES DESCRIPTIVES                                            *
*************************************************************************************************************;

proc tabulate data=dat.PROJECTE;
        class CASESTO SEX FUMA;
        tables (FUMA),SEX*CASESTO*(n*f=8.0 pctn<FUMA>='%'*f=6.2);
        title 'Diferències entre casos i controls segons variables explicatives.';
run;

proc tabulate data=dat.PROJECTE;
        class CASESTO SEX L_SCHOOL;
        tables (L_SCHOOL),SEX*CASESTO*(n*f=8.0 pctn<L_SCHOOL>='%'*f=6.2);
        title 'Diferències entre casos i controls segons variables explicatives.';
run;

proc tabulate data=dat.PROJECTE;
        class CASESTO SEX;var AGE_RECR QG07 QENER BMI;
        tables (AGE_RECR BMI QG07 QENER),SEX*CASESTO*(n*f=8.0 mean std);
        title 'Diferències entre casos i controls segons variables explicatives.';
run;

proc tabulate data=dat.PROJECTE missing;
        class CASESTO SEX CNTR_C;
        tables (CNTR_C all),SEX*CASESTO*(n*f=8.0);where CASESTO=1;
        title 'Distribució dels casos.';
run;



*** FINAL ***;
```

* CALIBPROJ.SAS

```
* Projecte de fi de carrera LCTE
* Programa en SAS que calibra les variables de dieta de l'EPIC
* (versió sense bootstrap)
* Cal executar PROJECTE.SAS
* Creació 22-03-2004 Modificació 29-09-04
* per Guillem Pera;


options pagesize=500 linesize=100 nonumber nodate;run;

libname DAT "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades";
libname library "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades";
run;


*** Model basat en el meeting IARC de Cambridge nov'03 ***;


* NOTA: És indiferent usar proc REG o proc GENMOD per estimar els valors predits. El segon procediment, però
és més flexible en quant a la sintaxi (a més de permetre canviar el link, la distribució,...).
També és anàleg construir la variable interacció que posar-la directament al model;

* El model utilitzat té les següents característiques:

1) model de regressió lineal normal amb efectes fixes.
2) l'intercept és específic per cada centre (uso els centres definits per CNTR_C).
3) la pendent (la lambda o coeficient de desatenuació) és específic per cada país (uso COUNTRY).
4) s'estimen els models separadament per SEX.
5) no s'usen transformacions.
6) s'usen les variables d'ajust proposades per la IARC: edat (AGE_RECR), estació del DQ (ESTADQn),
   pes (WEIGHT_C) i alçada (HEIGHT_C). Usar BMI, sol o acompanyat d'alguna de les dos variables anteriors no
   aporta millores interessants en el model, per tant continuo amb la proposta de la IARC.
7) es pondera per estació i dia del R24H per cada país i sexe.
8) no consumidors: s'imputa 0 directament als no consumidors (definits com a QGnn=0) ja que no he trobat cap
   altre punt de tall més atractiu. Si de cas es podria provar amb Q<=3 per tots els grups.
9) outliers: és molt difícil decidir a partir de quan el consum és excessiu. Els tests que incorpora el proc REG
   podrien ajudar, però sempre suposa l'eliminació de milers d'individus. Per tant, per ara, no faig res, excepte
   un anàlisi de sensibilitat en que s'eliminen els individus amb una ingesta en Q una mica superior del P99.
   Concretament els següents valors:
                                        HOMES                                                   DONES
        CARN              350 (550 per calibratge)    250 (350 per calibratge)
;


* Cal ordenar el dataset per fer un BY;
proc sort data=dat.PROJECTE;by SEX;run;

* ATENCIÓ!!!: fer WHERE amb OUT en el procediment GENMOD provoca l'eliminació dels individus que no
  compleixin l'ordre del WHERE. Per tant cal posar en OUT un altre dataset temporal;

*************************************************************************************************
*                              MODEL DE CALIBRATGE I AJUST DEL MODEL                            *
*************************************************************************************************;


*** MODEL DE CALIBRATGE:;

proc genmod data=dat.PROJECTE;
         class CNTR_C COUNTRY;
         model RG07=CNTR_C AGE_RECR  HEIGHT_C WEIGHT_C ESTADQ1 ESTADQ2 ESTADQ3 COUNTRY*QG07;by SEX;
         weight PONDER;
         output out=dat.KK07 predicted=CQG07;
         where QG07>0;* Només consumidors!;
run;


*** AFEGIR ELS ZEROS;

data dat.kk07;set dat.kk07;keep IDEPIC CQG07;run;

proc sort data=dat.PROJECTE;by IDEPIC;run;
proc sort data=dat.kk07;by IDEPIC;run;

data dat.PROJECTE;merge dat.PROJECTE dat.kk07;by IDEPIC;run;

data dat.PROJECTE;set dat.PROJECTE;
         if QG07=0 then CQG07=0;

         label CQG07="CARN (calibrada)";
run;
```

```
*** AJUST DEL MODEL;

* Aquesta instrucció serveix per calcular alguns tests de diagnòstic del model;
* Cal fer la regressió per separat en no haver-hi homes a França (1), Nàpols (25) i Utrecht (52);
proc reg data=dat.PROJECTE;* el centre de referència és 22 (Varese);
         model RG07=c21 c23 c24 c31--c51 c61--c92 AGE_RECR  HEIGHT_C WEIGHT_C ESTADQ1 ESTADQ2 ESTADQ3 c07_2--c07_9;
         weight PONDER;
         output out=dat.KK07bisM dffits=v1 h=v2 cookd=v3 press=v4 student=v5;
         * |valors| >2 indicarien problemes per dffits i student. Per h valors >14/13000 (homes) i >14/21000 (dones)
         indicarien problemes però sempre hi haurà centenars d'individus;
         where QG07>0 and SEX=1;* Homes;
run;

proc reg data=dat.PROJECTE;
         model RG07=c11--c21 c23--c92 AGE_RECR  HEIGHT_C WEIGHT_C ESTADQ1 ESTADQ2 ESTADQ3 c07_1--c07_9;
         weight PONDER;
         output out=dat.KK07bisF dffits=v1 h=v2 cookd=v3 press=v4 student=v5;
         where QG07>0 and SEX=2;* Dones;
run;

* Cal renombrar les variables;
data dat.kk07bisM;set dat.kk07bisM dat.kk07bisF;
         dfits07=v1;
         h07=v2;
         cook07=v3;
         press07=v4;
         stude07=v5;

         label    dfits07="Standard Influence on Predicted Value 07"
                  h07="Leverage 07"
                  cook07="Cook Influence 07"
                  press07="Residual without current obs 07"
                  stude07="Studentized Residual 07";

         keep IDEPIC dfits07 h07 cook07 press07 stude07;
run;

* Incorporació a la base de dades principal;
proc sort data=dat.kk07bisM;by IDEPIC;run;
proc sort data=dat.PROJECTE;by IDEPIC;run;

data dat.PROJECTE;merge dat.PROJECTE dat.kk07bisM;by IDEPIC;run;


*****************************************************************************************************
*                     REPETICIÓ DELS MODELS USANT TRANSFORMACIONS O EXCLUSIONS                      *
*****************************************************************************************************;


*** ARREL QUADRADA I LOGARITME;

proc sort data=dat.PROJECTE;by SEX;run;

proc glm data=dat.PROJECTE;* Arrel quadrada;
         class CNTR_C COUNTRY;
         model ARR07=CNTR_C AGE_RECR  HEIGHT_C WEIGHT_C ESTADQ1 ESTADQ2 ESTADQ3 COUNTRY*ARQ07 / solution;
         by SEX;
         weight PONDER;
         output out=dat.KK07 predicted=CAR07 dffits=v1 h=v2 cookd=v3 press=v4 student=v5;
         where QG07>0;
run;

proc glm data=dat.PROJECTE;* Logaritme;
         class CNTR_C COUNTRY;
         model LGR07=CNTR_C AGE_RECR  HEIGHT_C WEIGHT_C ESTADQ1 ESTADQ2 ESTADQ3 COUNTRY*LGQ07 / solution;
         by SEX;
         weight PONDER;
         output out=dat.KK07b predicted=CLG07 dffits=v1 h=v2 cookd=v3 press=v4 student=v5;
         where QG07>0;
run;

* Diagnòstic;
data dat.kk07;set dat.kk07;
         dfitsAR07=v1;
         hAR07=v2;
         cookAR07=v3;
         pressAR07=v4;
         studeAR07=v5;

         label    dfitsAR07="Standard Influence on Predicted Value AR07"
                  hAR07="Leverage AR07"
                  cookAR07="Cook Influence AR07"
                  pressAR07="Residual without current obs AR07"
                  studeAR07="Studentized Residual AR07"
```

```
                CAR07="CARN (arrel quadrada) calibrada";

        keep IDEPIC CAR07 dfitsAR07 hAR07 cookAR07 pressAR07 studeAR07;
run;

data dat.kk07b;set dat.kk07b;
        dfitsLG07=v1;
        hLG07=v2;
        cookLG07=v3;
        pressLG07=v4;
        studeLG07=v5;

        label   dfitsLG07="Standard Influence on Predicted Value LG07"
                hLG07="Leverage LG07"
                cookLG07="Cook Influence LG07"
                pressLG07="Residual without current obs LG07"
                studeLG07="Studentized Residual LG07"
                CLG07="CARN (log-2) calibrada";

        keep IDEPIC CLG07 dfitsLG07 hLG07 cookLG07 pressLG07 studeLG07;
run;


* Incorporació a la base de dades principal;
proc sort data=dat.kk07;by IDEPIC;run;
proc sort data=dat.kk07b;by IDEPIC;run;
proc sort data=dat.PROJECTE;by IDEPIC;run;

data dat.PROJECTE;merge dat.PROJECTE dat.kk07 dat.kk07b;by IDEPIC;run;

* Imputació dels zeros;
data dat.PROJECTE;set dat.PROJECTE;
        if QG07=0 then CAR07=0;
        if QG07=0 then CLG07=-3.321928095;* valor de log(0.1)/log(2);
run;


*** EXCLUSIÓ D'OUTLIERS;

proc sort data=dat.PROJECTE;by SEX;run;

proc glm data=dat.PROJECTE;
        class CNTR_C COUNTRY;
        model RG07=CNTR_C AGE_RECR  HEIGHT_C WEIGHT_C ESTADQ1 ESTADQ2 ESTADQ3 COUNTRY*QG07 / solution;
        by SEX;
        weight PONDER;
        output out=dat.KK07 predicted=COUT07 dffits=v1 h=v2 cookd=v3 press=v4 student=v5;
        where QG07>0 and ((SEX=1 and QG07<=350 and RG07<=550) or (SEX=2 and QG07<=250 and RG07<=350));
run;

* Diagnòstic;
data dat.kk07;set dat.kk07;
        dfitsOU07=v1;
        hOU07=v2;
        cookOU07=v3;
        pressOU07=v4;
        studeOU07=v5;

        label   dfitsOU07="Standard Influence on Predicted Value 07 (truncated)"
                hOU07="Leverage 07 (truncated)"
                cookOU07="Cook Influence 07 (truncated)"
                pressOU07="Residual without current obs 07 (truncated)"
                studeOU07="Studentized Residual 07 (truncated)"
                COUT07="CARN (truncada) calibrada";

        keep IDEPIC COUT07 dfitsOU07 hOU07 cookOU07 pressOU07 studeOU07;
run;


* Incorporació a la base de dades principal;
proc sort data=dat.kk07;by IDEPIC;run;
proc sort data=dat.PROJECTE;by IDEPIC;run;

data dat.PROJECTE;merge dat.PROJECTE dat.kk07;by IDEPIC;run;

* Imputació de zeros;
data dat.PROJECTE;set dat.PROJECTE;
        if QG07=0 then COUT07=0;
run;



*** RECODIFICACIÓ DELS CONSUMIDORS DE MENYS DE 3 GRAMS A NO-CONSUMIDORS;

proc sort data=dat.PROJECTE;by SEX;run;
```

```
proc glm data=dat.PROJECTE;
         class CNTR_C COUNTRY;
         model RG07=CNTR_C AGE_RECR  HEIGHT_C WEIGHT_C ESTADQ1 ESTADQ2 ESTADQ3 COUNTRY*QG07 / solution;
         by SEX;
         weight PONDER;
         output out=dat.KK07 predicted=Cnc07 dffits=v1 h=v2 cookd=v3 press=v4 student=v5;
         where QG07>=3;
run;

* Diagnòstic;
data dat.kk07;set dat.kk07;
         dfitsnc07=v1;
         hnc07=v2;
         cooknc07=v3;
         pressnc07=v4;
         studenc07=v5;

         label    dfitsnc07="Standard Influence on Predicted Value 07 (<3=non-consumer)"
                  hnc07="Leverage 07 (<3=non-consumer)"
                  cooknc07="Cook Influence 07 (<3=non-consumer)"
                  pressnc07="Residual without current obs 07 (<3=non-consumer)"
                  studenc07="Studentized Residual 07 (<3=non-consumer)"
                  Cnc07="CARN (<3=no consumidor) calibrada";

         keep IDEPIC Cnc07 dfitsnc07 hnc07 cooknc07 pressnc07 studenc07;
run;


* Incorporació a la base de dades principal;
proc sort data=dat.kk07;by IDEPIC;run;
proc sort data=dat.PROJECTE;by IDEPIC;run;

data dat.PROJECTE;merge dat.PROJECTE dat.kk07;by IDEPIC;run;

* Imputació de zeros;
data dat.PROJECTE;set dat.PROJECTE;
         if QG07<3 then Cnc07=0;
run;


*** AJUST PER ENERGIA;

proc sort data=dat.PROJECTE;by SEX;run;

proc glm data=dat.PROJECTE;
         class CNTR_C COUNTRY;
         model RG07=CNTR_C AGE_RECR  HEIGHT_C WEIGHT_C ESTADQ1 ESTADQ2 ESTADQ3 COUNTRY*QG07 QENER / solution;
         by SEX;
         weight PONDER;
         output out=dat.KK07 predicted=Cen07 dffits=v1 h=v2 cookd=v3 press=v4 student=v5;
         where QG07>=0;
run;

* Diagnòstic;
data dat.kk07;set dat.kk07;
         dfitsen07=v1;
         hen07=v2;
         cooken07=v3;
         pressen07=v4;
         studeen07=v5;

         label    dfitsen07="Standard Influence on Predicted Value 07 (energy adjusted)"
                  hen07="Leverage CARN (ajustat per energia)"
                  cooken07="Cook CARN (ajustat per energia)"
                  pressen07="Residual without current obs 07 (ajustat per energia)"
                  studeen07="Res Stud CARN (ajustat per energia)"
                  Cen07="CARN (ajustat per energia) calibrada";

         keep IDEPIC Cen07 dfitsen07 hen07 cooken07 pressen07 studeen07;
run;

* Incorporació a la base de dades principal;
proc sort data=dat.kk07;by IDEPIC;run;
proc sort data=dat.PROJECTE;by IDEPIC;run;

data dat.PROJECTE;merge dat.PROJECTE dat.kk07;by IDEPIC;run;

* Imputació de zeros;
data dat.PROJECTE;set dat.PROJECTE;
         if QG07=0 then Cen07=0;
run;
```

```
*** FINAL ***;
```

* BOOTSTRAP.SAS

```
*
* Projecte de fi de carrera LCTE
* Realitza procediment habitual i bootstrap per calibrar i calcular el HR corregit
* Creació: 12-10-2004       Modificació: 18-10-2004
* per Guillem Pera;



libname DAT "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades";
libname library "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades";
option  linesize=250 pagesize=500;
run;


*************************************************************************************************
*            CREACIÓ DEL DATASET "BOOT" QUE CONTÉ TOTES LES VARIABLES ÚTILS PEL BOOTSTRAP        *
*************************************************************************************************;

data dat.BOOT;set dat.PROJECTE;
         keep IDEPIC SEX COUNTRY CNTR_C
         c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92
         AGE_RECR AGEXIT CASESTO kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
         WEIGHT_C HEIGHT_C ESTADQ1 ESTADQ2 ESTADQ3 RG07 QG07 QENER RENER PONDER;
run;


*************************************************************************************************
*                                                                                               *
*     MACRO PER OBTENIR ELS COEFICIENTS I ERRORS ESTÀNDARD DE 3 TIPUS DE MODEL DE MALALTIA      *
*                  -        USANT ELS VALORS DEL QÜESTIONARI ORIGINAL (Q)                       *
*                  -        USANT VALORS PREDITS PEL CALIBRATGE (X)                             *
*                  -        USANT UN MÈTODE BOOTSTRAP                                           *
*                                                                                               *
*************************************************************************************************
*                                                                                               *
* NBOOT = nombre de repeticions                                                                 *
* SEED  = llavor aleatòria                                                                      *
* RRR   = variable del R24H (p.ex. RG07) en el model de calibratge                              *
* QQQ   = variable del qüestionari original (p.ex. QG07) en els models de calibratge i malaltia *
* NCON  = blanc    -anàlisis que inclouen tots els individus de l'estudi de calibratge          *
*          valor   -anàlisis en que els no consumidors, definits com els que consumeixen menys  *
*                  que aquest valor, són exclosos del model de calibratge (p.ex. 3 g/dia per    *
*                  definir no consumidors de carn (veure nota a sota))                          *
* DATAOUT = dataset amb els resultats de la macro                                               *
* STATUS    =      status respecte la malaltia (malalt (=1), censurat (=0))                     *
* BYVAR     =      variable usada per fer anàlisis separats en el model de malaltia             *
*                              (p.ex. SEX o blanc)                                              *
* STRATAVAR =      variable d'estratificació en el model de malaltia                            *
*                  (p.ex. COUNTRY o blanc)                                                      *
*                                                                                               *
*************************************************************************************************

*************************************************************************************************
*                                                                                               *
*        Es tenen en compte dos possibles formes de treballar amb els no consumidors:          *
*                                                                                               *
*        1)  Tots els individus s'inclouen en els models de calibratge i malaltia               *
*        2)  Els no consumidors, identificats a partir del qüestionari inicial (Q), s'exclouen  *
*        del model de calibratge i s'identifiquen amb una variable en el model de malaltia      *
*                                                                                               *
*        a) Ja que ingestes petites poden reflectir quantitats provinent de receptes (ingredients)  *
*        més que un consum reportat realment, valors menors que un cert valor arbitrari es      *
*        consideren zeros (p.ex. valors menors de 3 g/dia per consum de carn)                   *
*        b) Els valors < que aquest valor arbitrari predefinit s'exclouen del model de calibratge  *
*        c) Als no consumidors se'ls hi imputa un valor predit de zero                          *
*        d)  Al model de malaltia s'usa una variable indicadora                                 *
*        (0=no consumidors, 1=consumidors) junt amb el consum calibrat en una escala contínua   *
*                                                                                               *
*************************************************************************************************;
* Inici macro;
%macro BOOTSTRAP(NBOOT,SEED,RRR,QQQ,NCON,DATAOUT,STATUS,BYVAR,STRATAVAR);

%let PQQQ = P&QQQ;

* Nota: no usar apòstrofs en els comentaris dins de la macro. Dóna problemes;
* No separar les instruccions % que estiguin juntes;


*        SELECCIÓ DE LES VARIABLES NECESSÀRIES PER L,ANÀLISI. Es guarden en T_PM;
proc sort data=dat.BOOT
         out=T_PM (keep=&QQQ &RRR &STATUS &BYVAR &STRATAVAR
         PONDER SEX CNTR_C COUNTRY kkFUMA1 kkFUMA2 BMI kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
         WEIGHT_C HEIGHT_C ESTADQ1 ESTADQ2 ESTADQ3
```

```
         c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92
         AGE_RECR AGEXIT IDEPIC QENER RENER);
         by SEX COUNTRY IDEPIC;
         where &STATUS ne . and &QQQ ne .;
run;


*****************************************************
*         CALIBRATGE       (sense bootstrap)          *
*****************************************************;


***      MODELS DE CALIBRATGE ***;

* Model amb baixos consumidors definits com no consumidors (ambdós exclosos);
%if %Length(&NCON) ne 0 %then %do;
         proc glm data=T_PM noprint;
                  class COUNTRY CNTR_C;
                  model &RRR = &QQQ &QQQ*COUNTRY CNTR_C HEIGHT_C WEIGHT_C AGE_RECR ESTADQ1 ESTADQ2 ESTADQ3/ solution;
                  by   SEX;
                  weight PONDER;
                  output out=DBB1(keep=IDEPIC COUNTRY SEX &PQQQ) predicted=&PQQQ;
                  where &QQQ>&NCON;
         run ;
%end;
%else %do;
* Model amb baixos consumidors i no consumidors inclosos;
proc glm data=T_PM noprint;
         class COUNTRY CNTR_C;
    model &RRR = &QQQ &QQQ*COUNTRY CNTR_C HEIGHT_C WEIGHT_C AGE_RECR ESTADQ1 ESTADQ2 ESTADQ3/ solution;
         by   SEX;
         weight PONDER;
    output out=DBB2(keep=IDEPIC COUNTRY SEX &PQQQ) predicted=&PQQQ;
run ;
%end;


*        CREACIÓ DE LA VARIABLE INDICADORA PELS NO CONSUMIDORS (i fusió dels resultats amb la resta de variables);
data DB;
         %if %Length(&NCON) ne 0 %then %do;
                  merge DBB1 T_PM(in=A);
                  by SEX COUNTRY IDEPIC;
                  * INDICADOR ;
                  if &QQQ<=&NCON then CONS=0;else CONS=1;
         %end;
         %else %do;
                  merge DBB2 T_PM(in=A);
                  by SEX COUNTRY IDEPIC ;
         %end;

         if &PQQQ =. then &PQQQ =0 ;
         label &PQQQ="&QQQ -  VALORS PREDITS (CALIBRATS)";
run ;

* Cal ordenar si volem fer diferents models segons una variable;
%if %Length(&BYVAR) ne 0 %then %do;
         proc sort data=DB;by &BYVAR;run;
%end;

*** MODELS DE MALALTIA (models de Cox) ***;

* Models amb baixos consumidors definits com no consumidors;
%if %Length(&NCON) ne 0 %then %do;
*        USANT DADES CALIBRADES;
         proc phreg data=DB nosummary;
                  model (AGE_RECR,AGEXIT)*&STATUS(0)= &PQQQ CONS SEX
                  BMI QENER kkFUMA1 kkFUMA2 kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
                  c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92
                  /risklimits;
                  strata &STRATAVAR;
                  by &BYVAR;
                  ods output parameterestimates=CALIB (keep= &BYVAR VARIABLE ESTIMATE STDERR);
         run ;

*        USANT DADES ORIGINALS DEL QÜESTIONARI;
         proc phreg data=DB nosummary;
                  model (AGE_RECR,AGEXIT)*&STATUS(0)= &QQQ CONS SEX
                  BMI QENER kkFUMA1 kkFUMA2 kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
                  c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92
                  /risklimits;
                  strata &STRATAVAR;
                  by &BYVAR;
                  ods output parameterestimates=ORIG (keep= &BYVAR VARIABLE ESTIMATE STDERR);
         run;
%end;
```

```
%else %do;
* Models sense diferenciar consumidors de no consumidors;
*         USANT DADES CALIBRADES;
          proc phreg data=DB nosummary;
                    model (AGE_RECR,AGEXIT)*&STATUS(0)= &PQQQ SEX
                    BMI QENER kkFUMA1 kkFUMA2 kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
                    c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92
                    /risklimits;
                    strata &STRATAVAR;
                    by &BYVAR;
                    ods output parameterestimates=CALIB (keep= &BYVAR VARIABLE ESTIMATE STDERR);
          run ;

*         USANT DADES ORIGINALS DEL QÜESTIONARI;
          proc phreg data=DB nosummary;
                    model (AGE_RECR,AGEXIT)*&STATUS(0)= &QQQ SEX
                    BMI QENER kkFUMA1 kkFUMA2 kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
                    c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92
                    /risklimits;
                    strata &STRATAVAR;
                    by &BYVAR;
                    ods output parameterestimates=ORIG (keep= &BYVAR VARIABLE ESTIMATE STDERR);
          run;
%end;



************************************************
*                     BOOTSTRAP                *
************************************************;

*****************************************************************************************************************
* El procediment repetirà NBOOT cops el calibratge, usant el nombre total d,individus amb R24H, però no tots *
* els individus, sinó una mostra aleatòria (cada cop diferent) d,aquests amb repetició (o sigui, al calibrar *
* un individu pot estar repetit, i algun individu amb R24H pot estar exclós).                                 *
*                                                                                                                *
* D,aquesta forma, en cada iteració la variable calibrada (=estimada) serà diferent. El model de Cox usarà   *
* sempre els mateixos individus (tota la cohort amb tots els casos) però cada cop la variable calibrada serà *
* diferent. Aquesta variabilitat aportada per anar canviant la variable independent en cada iteració és la   *
* que s,afegirà a la variabilitat en si del paràmetre estimat.                                                *
*****************************************************************************************************************;

*         CREACIÓ D,UN DATASET ANOMENAT RECALL AMB NOMÉS ELS IDENTIFICADORS DELS INDIVIDUS DE L,ESTUDI DE CALIBRATGE;
data RECALL(drop=RENER NBREC);
          set T_PM END=NOTPM;
          retain NBREC 0;
          if RENER ne . then do;
                    output;
                    NBREC=NBREC+1;
          end;
          if NOTPM eq 1 then call symput("NBREC",NBREC);
run ;

* Inici del bucle principal;
%do IBOOT1 = 1 %to &NBOOT;
*         GENERO UN NOU DATASET ALEATORI, T_RAND, AMB EL NOMBRE D,OBSERVACIONS DE RECALL;
          data T_RAND (drop=I U SEED);
                    retain SEED &SEED;
                    array NOBS{&NBREC} _TEMPORARY_ (&NBREC* 0);
                    do I = 1 to &NBREC;
                              call ranuni (SEED,U);
                              NOBS{I} = int(U * &NBREC ) + 1;
                    end;
                    do I = 1 to &NBREC;
                              IND = NOBS{I} ;
                              set RECALL POINT=IND ;
                              output;
                    end;
                    stop;
                    call symput("SEED",put(SEED,12.));
          run ;

* Afegeixo variables necessàries a fitxer aleatori;
          data BOOTS;
                    set T_RAND(in=A) T_PM(in=B);
                    if B eq 1 then &RRR = .;* Per comptes d,esborrar individus faig missing la variable d,interès;
          run ;

          proc sort data=BOOTS;
                    by SEX COUNTRY IDEPIC ;
          run ;


*** MODELS DE CALIBRATGE ***;
```

```
* Model amb baixos consumidors definits com no consumidors (ambdós exclosos);
        %if %Length(&NCON) ne 0 %then %do;
                proc glm data=BOOTS noprint;
                        class COUNTRY CNTR_C;
                model &RRR = &QQQ &QQQ*COUNTRY CNTR_C HEIGHT_C WEIGHT_C AGE_RECR ESTADQ1 ESTADQ2 ESTADQ3/ solution;
                        by   SEX;
                        weight PONDER;
                output out=DB1(keep=IDEPIC COUNTRY SEX &PQQQ) predicted = &PQQQ;
                        where &QQQ>&NCON ;
                run ;
        %end;
        %else %do ;
* Model amb baixos consumidors i no consumidors inclosos;
                proc glm data=BOOTS noprint;
                        class COUNTRY CNTR_C;
                model &RRR = &QQQ &QQQ*COUNTRY CNTR_C HEIGHT_C WEIGHT_C AGE_RECR ESTADQ1 ESTADQ2 ESTADQ3/ solution;
                        by   SEX;
                        weight PONDER;
                        output out=DB2(keep=IDEPIC COUNTRY SEX &PQQQ) predicted = &PQQQ;
                run;
        %end;


*       CREACIÓ DE LA VARIABLE INDICADORA DE NO CONSUMIDOR (i fusió dels resultats amb la resta de variables);
        data DB;
                %if %Length(&NCON) ne 0 %then %do;
                        merge DB1 T_PM(in=A) END=NODB;
                        by SEX COUNTRY IDEPIC;
                        * INDICADOR ;
                        if &QQQ<=&NCON then CONS=0;else CONS=1;
                        label CONS='CONSUMIDOR DE ...';
                %end;
                %else %do;
                        merge DB2 T_PM(in=A) END=NODB;
                        by SEX COUNTRY IDEPIC ;
                %end;

                if A eq 0 or &RRR ne . then delete;* DB tindrà tota la cohort per fer el model de Cox;
                if &PQQQ =. then &PQQQ =0;
                label &PQQQ='CONSUM PREDIT (CALIBRAT)';
        run ;

        %if %Length(&BYVAR) ne 0 %then %do;
                proc sort data=DB;* Cal ordenar si volem fer diferents models segons una variable;
                        by &BYVAR;
                run;
        %end;

        ods exclude all;

*** MODELS DE MALALTIA (models de Cox) ***;

*       USANT DADES CALIBRADES;

* Models amb baixos consumidors definits com no consumidors;
        %if %Length(&NCON) ne 0 %then %do;
                proc phreg data=DB;
                        model (AGE_RECR,AGEXIT)*&STATUS(0)= &PQQQ CONS SEX
                        BMI QENER kkFUMA1 kkFUMA2 kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
                        c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82
                        c91 c92/risklimits;
                        strata &STRATAVAR ;
                        by &BYVAR ;
                        ods output parameterestimates=TMP (where= (VARIABLE="&PQQQ"));
                run ;

                proc append BASE=&DATAOUT data=TMP(keep = &BYVAR ESTIMATE STDERR);
                run;* Afegeixo resultats al fitxer especificat a la instrucció BOOTSTRAP;

        %end;
        %else %do;
* Models sense diferenciar consumidors de no consumidors;
                proc phreg data=DB;
                        model (AGE_RECR,AGEXIT)*&STATUS(0)= &PQQQ SEX
                        BMI QENER kkFUMA1 kkFUMA2 kkSCHOO1 kkSCHOO2 kkSCHOO3 kkSCHOO4 kkSCHOO5
                        c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82
                        c91 c92/risklimits;
                        strata &STRATAVAR;
                        by &BYVAR;
                        ods output parameterestimates=TMP (where= (VARIABLE="&PQQQ"));
                run;

                proc append BASE=&DATAOUT data=TMP(keep = &BYVAR ESTIMATE STDERR);
                run ;* Afegeixo resultats al fitxer especificat a la instrucció BOOTSTRAP;
```

```
        %end;

        proc datasets library=WORK nodetails;* Esborro fitxers;
                delete T_RAND BOOTS DB1 DB2 DB TMP;
        quit;

%end;* Final del bucle principal;


*       OBTENCIÓ DELS COEFICIENTS I DELS ERRORS ESTÀNDARD DEL MODEL DE MALALTIA MITJANÇANT BOOTSTRAP;

data &DATAOUT;set &DATAOUT;
        VAR = STDERR**2;* Variància;
run;

ods select all;

proc means data=&DATAOUT n mean std;
        var ESTIMATE VAR;
        class &BYVAR;
        ods output summary=BO (keep = &BYVAR ESTIMATE_STDDEV VAR_MEAN);
run;* En BO es guarden les dades resumides del fitxer de resultats;
* ESTIMATE_STDDEV és la desviació estàndard de les NBOOT beta estimades, o sigui, la variabilitat aportada
  pel bootstrap;
* VAR_MEAN és la mitjana de les NBOOT variàncies estimades de beta, o sigui, la variabilitat de beta en si;

* Preparació dels fitxers per l,output;
data ORIG;set ORIG;where VARIABLE="&QQQ";run;
data CALIB;set CALIB;where VARIABLE="&PQQQ";run;

%if %Length(&BYVAR) ne 0 %then %do;
        proc sort data=ORIG;by &BYVAR; run;
        proc sort data=CALIB;by &BYVAR; run;
        proc sort data=BO;by &BYVAR; run;
%end;


*       CÀLCUL DE L,ERROR ESTÀNDARD CORREGIT;
data SE;
        merge ORIG (keep = VARIABLE &BYVAR ESTIMATE STDERR )
              CALIB(keep = VARIABLE &BYVAR ESTIMATE STDERR rename=(ESTIMATE=PEST STDERR = PSTD))
              BO   (keep = &BYVAR ESTIMATE_STDDEV VAR_MEAN);
        by &BYVAR ;
        BOOTSTD = ESTIMATE_STDDEV;
        SE_CORR = sqrt(VAR_MEAN + ESTIMATE_STDDEV**2);
        * SE_CORR és el que busquem: l,error estàndard corregit, format per la suma de la variància mitjana i
          la variància de calcular l,estimador de beta NBOOT cops;
        TITLE  - &STATUS - &QQQ -;
        label    ESTIMATE = 'beta (dades originals)'
                 STDERR    = 'SE(beta) (dades originals)'
                 PEST      = 'beta calibrada (corregida)'
                 PSTD      = 'SE(beta calibrada (corregida))'
                 BOOTSTD  = 'SE (bootstrap) (variabilitat aportada pel bootstrap)'
                 SE_CORR  = 'SE CORREGIT (VIA BOOTSTRAP) DE LA BETA CALIBRADA. Variabilitat total.';
run;


*       OUTPUT ;
%if %Length(&BYVAR) ne 0 %then %do;
        proc print data=SE label;
                 var &BYVAR  ESTIMATE STDERR PEST PSTD BOOTSTD SE_CORR;
                 where &BYVAR ne .;
        run;
%end;
%else %do ;
        proc print data=SE label;
                 var ESTIMATE STDERR PEST PSTD BOOTSTD SE_CORR;
        run;
%end;

%mend;



* Final de macro;



/*** Exemples d'utilització ***

*per sexe amb consum<=3 considerat no consumidor;
%BOOTSTRAP (300,-19892003,RG07,QG07,3 ,dat.resultats1, CASESTO,SEX, );
```

```
*estratificat per país amb consum<=3 considerat no consumidor;
%BOOTSTRAP (300,-19892003,RG07,QG07,3 ,dat.resultats2, CASESTO,  ,COUNTRY);

*per sexe, estratificat per país, incloent no consumidors en el calibratge;
%BOOTSTRAP (300,-19892003,RG07,QG07, ,dat.resultats3, CASESTO,SEX,COUNTRY );

*anàlisi simple;
%BOOTSTRAP (300,-19892003,RG07,QG07, ,dat.resultats4, CASESTO,  , );  */

** NOTA IMPOPORTANT **
El fitxer de resultats va creixent en cada nova execució de BOOTSTRAP. És aconsellable canviar el nom
en cada execució o esborrar el fitxer antic si no es vol que els resultats es vagin acumulant;



* Instrucció emprada pel projecte: ;

%BOOTSTRAP (300,-19892003,RG07,QG07,0.001,dat.BOOTSTRAP,CASESTO, ,COUNTRY);


*** FINAL ***;
```

* BOXCOX.DO

```
* Projecte de fi de carrera LCTE
* Programa en Stata
* pel càlcul de la transformació més adient de les dades per assolir normalitat/homocedasticitat usant Box- Cox
* i per veure la bondat de l'ajust del model de Cox
* per Guillem Pera


*********************************************************************************************************
*                                        TRANSFORMACIÓ DE BOX-COX                                       *
*********************************************************************************************************


clear
set mem 350m

log using "C:\GP\Metodologia\Calibratge\Projecte UPC\boxcoxmen.log"

use "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades\boxcoxmen.dta", clear

************* HOMES *************

* Cal sumar una constant per no tenir zeros
gen rg07b=rg07+1
gen qg07b=qg07+1

* Càlcul de la millor transformació per QFA/HD
boxcox qg07b if country=="2"
boxcox qg07b if country=="3"
boxcox qg07b if country=="4"
boxcox qg07b if country=="5"
boxcox qg07b if country=="6"
boxcox qg07b if country=="7"
boxcox qg07b if country=="8"
boxcox qg07b if country=="9"
boxcox qg07b if rg07b>0

* Càlcul de la millor transformació per R24H
boxcox rg07b if country=="2"
boxcox rg07b if country=="3"
boxcox rg07b if country=="4"
boxcox rg07b if country=="5"
boxcox rg07b if country=="6"
boxcox rg07b if country=="7"
boxcox rg07b if country=="8"
boxcox rg07b if country=="9"
boxcox rg07b


save "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades\boxcoxmen.dta", replace
log close


************* DONES *************

log using "C:\GP\Metodologia\Calibratge\Projecte UPC\boxcoxwomen.log"

use "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades\boxcoxwomen.dta", clear

* Cal sumar una constant per no tenir zeros
gen rg07b=rg07+1
gen qg07b=qg07+1

* Càlcul de la millor transformació per QFA/HD
boxcox qg07b if country=="2"
boxcox qg07b if country=="3"
boxcox qg07b if country=="4"
boxcox qg07b if country=="5"
boxcox qg07b if country=="6"
boxcox qg07b if country=="7"
boxcox qg07b if country=="8"
boxcox qg07b if country=="9"
boxcox qg07b if rg07b>0

* Càlcul de la millor transformació per R24H
boxcox rg07b if country=="2"
boxcox rg07b if country=="3"
boxcox rg07b if country=="4"
boxcox rg07b if country=="5"
boxcox rg07b if country=="6"
boxcox rg07b if country=="7"
boxcox rg07b if country=="8"
boxcox rg07b if country=="9"
```

```
boxcox rg07b


save "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades\boxcoxwomen.dta", replace
save "C:\guillem\ico\Calibratge\Projecte UPC\Dades\boxcoxwomen.dta", replace

log close


*************************************************************************************************
*                                      AVALUACIÓ DEL MODEL DE COX                                 *
*************************************************************************************************

use "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades\ajcox10.dta", clear

* Definició del set de dades per supervivència
stset agexit, enter(age_recr) fail(casesto)

* delimit serveix per trencar línies llargues
#delimit ;
stcox kkfuma1 kkfuma2 bmi kkschoo1 kkschoo2 kkschoo3 kkschoo4 kkschoo5 qener cqg07c z07 sex
c11 c12 c13 c14 c21 c22 c23 c24 c25 c31 c32 c33 c34 c35 c41 c42 c43 c51 c52 c61 c71 c72 c81 c82 c91 c92,
 strata(country) mgale(martin) schoenfeld(sch*) scaledsch(sca*);
#delimit cr

* Testa assumpció de proporcionalitat
stphtest, detail
predict double coxsnell, csnell

* Gràfica de residus de martingala respecte edat
lowess martin age_recr,mean noweight

* Gràfica de Cox-Snell
stset coxsnell, failure (casesto)
sts generate km=s
generate double H=-ln(km)
line H coxsnell coxsnell,sort ytitle(" ") clstyle(. refline) legend(nodraw)

save "C:\GP\Metodologia\Calibratge\Projecte UPC\Dades\ajcox10.dta", replace

*** FINAL ***;
```